MOLECULES FOR DISEASE DETECTION AND TREATMENT

TECHNICAL FIELD

The present invention relates to molecules for disease detection and treatment and to the use of these sequences in the diagnosis, study, prevention, and treatment of diseases associated with disease detection and treatment molecules.

BACKGROUND OF THE INVENTION

The human genome is comprised of thousands of genes, many encoding gene products that function in the maintenance and growth of the various cells and tissues in the body. Aberrant expression or mutations in these genes and their products is the cause of, or is associated with, a variety of human diseases such as cancer and other cell proliferative disorders. The identification of these genes and their products is the basis of an ever-expanding effort to find markers for early detection of diseases, and targets for their prevention and treatment.

For example, cancer represents a type of cell proliferative disorder that affects nearly every tissue in the body. A wide variety of molecules, either aberrantly expressed or mutated, can be the cause of, or involved with, various cancers because tissue growth involves complex and ordered patterns of cell proliferation, cell differentiation, and apoptosis. Cell proliferation must be regulated to maintain both the number of cells and their spatial organization. This regulation depends upon the appropriate expression of proteins which control cell cycle progression in response to extracellular signals such as growth factors and other mitogens, and intracellular cues such as DNA damage or nutrient starvation. Molecules which directly or indirectly modulate cell cycle progression fall into several categories, including growth factors and their receptors, second messenger and signal transduction proteins, oncogene products, tumor-suppressor proteins, and mitosis-promoting factors. Aberrant expression or mutations in any of these gene products can result in cell proliferative disorders such as cancer. Oncogenes are genes generally derived from normal genes that, through abnormal expression or mutation, can effect the transformation of a normal cell to a malignant one (oncogenesis). Oncoproteins, encoded by oncogenes, can affect cell proliferation in a variety of ways and include growth factors, growth factor receptors, intracellular signal transducers, nuclear transcription factors, and cell-cycle control proteins. In contrast, tumor-suppressor genes are involved in inhibiting cell proliferation. Mutations which cause reduced or loss of function in tumor-suppressor genes result in aberrant cell proliferation and cancer. Thus a wide variety of genes and their products have been found that are associated with cell proliferative disorders such as cancer, but many more may exist that are yet to be discovered.

DNA-based arrays can provide a simple way to explore the expression of a single polymorphic gene or a large number of genes. When the expression of a single gene is explored,

DNA-based arrays are employed to detect the expression of specific gene variants. For example, a p53 tumor suppressor gene array is used to determine whether individuals are carrying mutations that predispose them to cancer. A cytochrome p450 gene array is useful to determine whether individuals have one of a number of specific mutations that could result in increased drug metabolism, drug resistance or drug toxicity.

DNA-based array technology is especially relevant for the rapid screening of expression of a large number of genes. There is a growing awareness that gene expression is affected in a global fashion. A genetic predisposition, disease or therapeutic treatment may affect, directly or indirectly, the expression of a large number of genes. In some cases the interactions may be expected, such as when the genes are part of the same signaling pathway. In other cases, such as when the genes participate in separate signaling pathways, the interactions may be totally unexpected. Therefore, DNA-based arrays can be used to investigate how genetic predisposition, disease, or therapeutic treatment affects the expression of a large number of genes.

The discovery of new molecules for disease detection and treatment satisfies a need in the art by providing new compositions which are useful in the diagnosis, study, prevention, and treatment of diseases.

SUMMARY OF THE INVENTION

The present invention relates to human polynucleotides encoding molecules for disease detection and treatment (mddt) as presented in the Sequence Listing. Some of the mddt uniquely identify genes encoding structural, functional, and regulatory molecules for disease detection and treatment.

The invention provides an isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). In one alternative, the polynucleotide comprises a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14. In another alternative, the polynucleotide comprises at least 60 contiguous nucleotides of a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The invention further provides a composition for

the detection of expression of disease detection and treatment molecule polynucleotides comprising at least one isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d); and a detectable label.

The invention also provides a method for detecting a target polynucleotide in a sample, said target polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The method comprises a) hybridizing the sample with a probe comprising at least 20 contiguous nucleotides comprising a sequence complementary to said target polynucleotide in the sample, and which probe specifically hybridizes to said target polynucleotide, under conditions whereby a hybridization complex is formed between said probe and said target polynucleotide, and b) detecting the presence or absence of said hybridization complex, and, optionally, if present, the amount thereof. In one alternative, the probe comprises at least 30 contiguous nucleotides. In another alternative, the probe comprises at least 60 contiguous nucleotides.

The invention further provides a recombinant polynucleotide comprising a promoter sequence operably linked to an isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). In one alternative, the invention provides a cell transformed with the recombinant polynucleotide. In another alternative, the invention provides a transgenic organism comprising the recombinant polynucleotide. In a further alternative, the invention provides a method for producing a disease detection and treatment molecule polypeptide, the method comprising a) culturing a cell under conditions suitable for expression of the disease detection and treatment molecule polypeptide, wherein said cell is transformed with the recombinant polynucleotide, and b) recovering the disease detection and treatment molecule polypeptide so expressed.

The invention also provides a purified disease detection and treatment molecule polypeptide (MDDT) encoded by at least one polynucleotide comprising a polynucleotide sequence selected from

the group consisting of SEQ ID NO:1-14. Additionally, the invention provides an isolated antibody which specifically binds to the disease detection and treatment molecule polypeptide. The invention further provides a method of identifying a test compound which specifically binds to the disease detection and treatment molecule polypeptide, the method comprising the steps of a) providing a test compound; b) combining the disease detection and treatment molecule polypeptide with the test compound for a sufficient time and under suitable conditions for binding; and c) detecting binding of the disease detection and treatment molecule polypeptide to the test compound, thereby identifying the test compound which specifically binds the disease detection and treatment molecule polypeptide.

The invention further provides a microarray wherein at least one element of the microarray is an isolated polynucleotide comprising at least 60 contiguous nucleotides of a polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The invention also provides a method of using the microarray for generating a transcript image of a sample which contains polynucleotides. The method comprises a) labeling the polynucleotides of the sample, b) contacting the elements of the microarray with the labeled polynucleotides of the sample under conditions suitable for the formation of a hybridization complex, and c) quantifying the expression of the polynucleotides in the sample.

Additionally, the invention provides a method for screening a compound for effectiveness in altering expression of a target polynucleotide, wherein said target polynucleotide comprises a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b); and e) an RNA equivalent of a) through d). The method comprises a) exposing a sample comprising the target polynucleotide to a compound, and b) detecting altered expression of the target polynucleotide.

The invention further provides a method for detecting a target polynucleotide in a sample for toxicity testing of a compound, said target polynucleotide comprising a polynucleotide sequence selected from the group consisting of a) a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence selected from the group consisting of SEQ ID NO:1-14; c) a polynucleotide sequence complementary to a); d) a polynucleotide sequence complementary to b);

and e) an RNA equivalent of a) through d). The method comprises a) hybridizing the sample with a probe comprising at least 20 contiguous nucleotides comprising a sequence complementary to said target polynucleotide in the sample, and which probe specifically hybridizes to said target polynucleotide, under conditions whereby a hybridization complex is formed between said probe and said target polynucleotide, b) detecting the presence or absence of said hybridization complex, and, optionally, if present, the amount thereof, and c) comparing the presence, absence or amount of said target polynucleotide in a first biological sample and a second biological sample, wherein said first biological sample has been contacted with said compound, and said second sample is a control, whereby a change in presence, absence or amount of said target polynucleotide in said first sample, as compared with said second sample, is indicative of toxic response to said compound.

DESCRIPTION OF THE TABLES

Table 1 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with their GenBank hits (GI Numbers), probability scores, and functional annotations corresponding to the GenBank hits.

Table 2 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with polynucleotide segments of each template sequence as defined by the indicated "start" and "stop" nucleotide positions. The reading frames of the polynucleotide segments and the Pfam hits, Pfam descriptions, and E-values corresponding to the polypeptide domains encoded by the polynucleotide segments are indicated.

Table 3 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with polynucleotide segments of each template sequence as defined by the indicated "start" and "stop" nucleotide positions. The reading frames of the polynucleotide segments are shown, and the polypeptides encoded by the polynucleotide segments constitute either signal peptide (SP) or transmembrane (TM) domains, as indicated.

Table 4 shows the sequence identification numbers (SEQ ID NO:s) and template identification numbers (template IDs) corresponding to the polynucleotides of the present invention, along with component sequence identification numbers (component IDs) corresponding to each template. The component sequences, which were used to assemble the template sequences, are defined by the indicated "start" and "stop" nucleotide positions along each template.

Table 5 summarizes the bioinformatics tools which are useful for analysis of the polynucleotides of the present invention. The first column of Table 5 lists analytical tools, programs,

and algorithms, the second column provides brief descriptions thereof, the third column presents appropriate references, all of which are incorporated by reference herein in their entirety, and the fourth column presents, where applicable, the scores, probability values, and other parameters used to evaluate the strength of a match between two sequences (the higher the score, the greater the homology between two sequences).

DETAILED DESCRIPTION OF THE INVENTION

Before the nucleic acid sequences and methods are presented, it is to be understood that this invention is not limited to the particular machines, methods, and materials described. Although particular embodiments are described, machines, methods, and materials similar or equivalent to these embodiments may be used to practice the invention. The preferred machines, methods, and materials set forth are not intended to limit the scope of the invention which is limited only by the appended claims.

The singular forms "a", "an", and "the" include plural reference unless the context clearly dictates otherwise. All technical and scientific terms have the meanings commonly understood by one of ordinary skill in the art. All publications are incorporated by reference for the purpose of describing and disclosing the cell lines, vectors, and methodologies which are presented and which might be used in connection with the invention. Nothing in the specification is to be construed as an admission that the invention is not entitled to antedate such disclosure by virtue of prior invention.

Definitions

As used herein, the lower case "mddt" refers to a nucleic acid sequence, while the upper case "MDDT" refers to an amino acid sequence encoded by mddt. A "full-length" mddt refers to a nucleic acid sequence containing the entire coding region of a gene endogenously expressed in human tissue.

"Adjuvants" are materials such as Freund's adjuvant, mineral gels (aluminum hydroxide), and surface active substances (lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanin, and dinitrophenol) which may be administered to increase a host's immunological response.

"Allele" refers to an alternative form of a nucleic acid sequence. Alleles result from a "mutation," a change or an alternative reading of the genetic code. Any given gene may have none, one, or many allelic forms. Mutations which give rise to alleles include deletions, additions, or substitutions of nucleotides. Each of these changes may occur alone, or in combination with the others, one or more times in a given nucleic acid sequence. The present invention encompasses allelic mddt.

"Amino acid sequence" refers to a peptide, a polypeptide, or a protein of either natural or

synthetic origin. The amino acid sequence is not limited to the complete, endogenous amino acid sequence and may be a fragment, epitope, variant, or derivative of a protein expressed by a nucleic acid sequence.

"Amplification" refers to the production of additional copies of a sequence and is carried out using polymerase chain reaction (PCR) technologies well known in the art.

"Antibody" refers to intact molecules as well as to fragments thereof, such as Fab, $F(ab')_2$, and Fv fragments, which are capable of binding the epitopic determinant. Antibodies that bind MDDT polypeptides can be prepared using intact polypeptides or using fragments containing small peptides of interest as the immunizing antigen. The polypeptide or peptide used to immunize an animal (e.g., a mouse, a rat, or a rabbit) can be derived from the translation of RNA, or synthesized chemically, and can be conjugated to a carrier protein if desired. Commonly used carriers that are chemically coupled to peptides include bovine serum albumin, thyroglobulin, and keyhole limpet hemocyanin (KLH). The coupled peptide is then used to immunize the animal.

"Antisense sequence" refers to a sequence capable of specifically hybridizing to a target sequence. The antisense sequence may include DNA, RNA, or any nucleic acid mimic or analog such as peptide nucleic acid (PNA); oligonucleotides having modified backbone linkages such as phosphorothioates, methylphosphonates, or benzylphosphonates; oligonucleotides having modified sugar groups such as 2'-methoxyethyl sugars or 2'-methoxyethoxy sugars; or oligonucleotides having modified bases such as 5-methyl cytosine, 2'-deoxyuracil, or 7-deaza-2'-deoxyguanosine.

"Antisense sequence" refers to a sequence capable of specifically hybridizing to a target sequence. The antisense sequence can be DNA, RNA, or any nucleic acid mimic or analog.

"Antisense technology" refers to any technology which relies on the specific hybridization of an antisense sequence to a target sequence.

A "bin" is a portion of computer memory space used by a computer program for storage of data, and bounded in such a manner that data stored in a bin may be retrieved by the program.

"Biologically active" refers to an amino acid sequence having a structural, regulatory, or biochemical function of a naturally occurring amino acid sequence.

"Clone joining" is a process for combining gene bins based upon the bins' containing sequence information from the same clone. The sequences may assemble into a primary gene transcript as well as one or more splice variants.

"Complementary" describes the relationship between two single-stranded nucleic acid sequences that anneal by base-pairing (5'-A-G-T-3' pairs with its complement 3'-T-C-A-5').

A "component sequence" is a nucleic acid sequence selected by a computer program such as PHRED and used to assemble a consensus or template sequence from one or more component sequences.

A "consensus sequence" or "template sequence" is a nucleic acid sequence which has been assembled from overlapping sequences, using a computer program for fragment assembly such as the GELVIEW fragment assembly system (Genetics Computer Group (GCG), Madison WI) or using a relational database management system (RDMS).

"Conservative amino acid substitutions" are those substitutions that, when made, least interfere with the properties of the original protein, i.e., the structure and especially the function of the protein is conserved and not significantly changed by such substitutions. The table below shows amino acids which may be substituted for an original amino acid in a protein and which are regarded as conservative substitutions.

Original Residue	Conservative Substitution
Ala	Gly, Ser
Arg	His, Lys
Asn	Asp, Gln, His
Asp	Asn, Glu
Cys	Ala, Ser
Gln	Asn, Glu, His
Glu	Asp, Gln, His
Gly	Ala
His	Asn, Arg, Gln, Glu
Ile	Leu, Val
Leu	Ile, Val
Lys	Arg, Gln, Glu
Met	Leu, Ile
Phe	His, Met, Leu, Trp, Tyr
Ser	Cys, Thr
Thr	Ser, Val
Trp	Phe, Tyr
Tyr	His, Phe, Trp
Val	Ile, Leu, Thr

Conservative substitutions generally maintain (a) the structure of the polypeptide backbone in the area of the substitution, for example, as a beta sheet or alpha helical conformation, (b) the charge or hydrophobicity of the molecule at the target site, or (c) the bulk of the side chain.

"Deletion" refers to a change in either a nucleic or amino acid sequence in which at least one nucleotide or amino acid residue, respectively, is absent.

"Derivative" refers to the chemical modification of a nucleic acid sequence, such as by replacement of hydrogen by an alkyl, acyl, amino, hydroxyl, or other group.

The terms "element" and "array element" refer to a polynucleotide, polypeptide, or other chemical compound having a unique and defined position on a microarray.

"E-value" refers to the statistical probability that a match between two sequences occurred by chance.

A "fragment" is a unique portion of mddt or MDDT which is identical in sequence to but shorter in length than the parent sequence. A fragment may comprise up to the entire length of the defined sequence, minus one nucleotide/amino acid residue. For example, a fragment may comprise from 10 to 1000 contiguous amino acid residues or nucleotides. A fragment used as a probe, primer, antigen, therapeutic molecule, or for other purposes, may be at least 5, 10, 15, 16, 20, 25, 30, 40, 50, 60, 75, 100, 150, 250 or at least 500 contiguous amino acid residues or nucleotides in length. Fragments may be preferentially selected from certain regions of a molecule. For example, a polypeptide fragment may comprise a certain length of contiguous amino acids selected from the first 250 or 500 amino acids (or first 25% or 50%) of a polypeptide as shown in a certain defined sequence. Clearly these lengths are exemplary, and any length that is supported by the specification, including the Sequence Listing and the figures, may be encompassed by the present embodiments.

A fragment of mddt comprises a region of unique polynucleotide sequence that specifically identifies mddt, for example, as distinct from any other sequence in the same genome. A fragment of mddt is useful, for example, in hybridization and amplification technologies and in analogous methods that distinguish mddt from related polynucleotide sequences. The precise length of a fragment of mddt and the region of mddt to which the fragment corresponds are routinely determinable by one of ordinary skill in the art based on the intended purpose for the fragment.

A fragment of MDDT is encoded by a fragment of mddt. A fragment of MDDT comprises a region of unique amino acid sequence that specifically identifies MDDT. For example, a fragment of MDDT is useful as an immunogenic peptide for the development of antibodies that specifically recognize MDDT. The precise length of a fragment of MDDT and the region of MDDT to which the fragment corresponds are routinely determinable by one of ordinary skill in the art based on the intended purpose for the fragment.

A "full length" nucleotide sequence is one containing at least a start site for translation to a protein sequence, followed by an open reading frame and a stop site, and encoding a "full length" polypeptide.

"Hit" refers to a sequence whose annotation will be used to describe a given template. Criteria for selecting the top hit are as follows: if the template has one or more exact nucleic acid matches, the top hit is the exact match with highest percent identity. If the template has no exact matches but has significant protein hits, the top hit is the protein hit with the lowest E-value. If the template has no significant protein hits, but does have significant non-exact nucleotide hits, the top hit is the nucleotide hit with the lowest E-value.

"Homology" refers to sequence similarity either between a reference nucleic acid sequence and at least a fragment of an mddt or between a reference amino acid sequence and a fragment of an MDDT.

"Hybridization" refers to the process by which a strand of nucleotides anneals with a complementary strand through base pairing. Specific hybridization is an indication that two nucleic acid sequences share a high degree of identity. Specific hybridization complexes form under defined annealing conditions, and remain hybridized after the "washing" step. The defined hybridization conditions include the annealing conditions and the washing step(s), the latter of which is particularly important in determining the stringency of the hybridization process, with more stringent conditions allowing less non-specific binding, i.e., binding between pairs of nucleic acid probes that are not perfectly matched. Permissive conditions for annealing of nucleic acid sequences are routinely determinable and may be consistent among hybridization experiments, whereas wash conditions may be varied among experiments to achieve the desired stringency.

Generally, stringency of hybridization is expressed with reference to the temperature under which the wash step is carried out. Generally, such wash temperatures are selected to be about 5°C to 20°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. An equation for calculating T_m and conditions for nucleic acid hybridization is well known and can be found in Sambrook et al., 1989, Molecular Cloning: A Laboratory Manual, 2nd ed., vol. 1-3, Cold Spring Harbor Press, Plainview NY; specifically see volume 2, chapter 9.

High stringency conditions for hybridization between polynucleotides of the present invention include wash conditions of 68°C in the presence of about 0.2 x SSC and about 0.1% SDS, for 1 hour. Alternatively, temperatures of about 65°C, 60°C, or 55°C may be used. SSC concentration may be varied from about 0.2 to 2 x SSC, with SDS being present at about 0.1%. Typically, blocking reagents are used to block non-specific hybridization. Such blocking reagents include, for instance, denatured salmon sperm DNA at about 100-200 µg/ml. Useful variations on these conditions will be readily apparent to those skilled in the art. Hybridization, particularly under high stringency conditions, may be suggestive of evolutionary similarity between the nucleotides. Such similarity is strongly indicative of a similar role for the nucleotides and their resultant proteins.

Other parameters, such as temperature, salt concentration, and detergent concentration may be varied to achieve the desired stringency. Denaturants, such as formamide at a concentration of about 35-50% v/v, may also be used under particular circumstances, such as RNA:DNA hybridizations. Appropriate hybridization conditions are routinely determinable by one of ordinary skill in the art.

"Immunogenic" describes the potential for a natural, recombinant, or synthetic peptide, epitope, polypeptide, or protein to induce antibody production in appropriate animals, cells, or cell lines.

"Insertion" or "addition" refers to a change in either a nucleic or amino acid sequence in which at least one nucleotide or residue, respectively, is added to the sequence.

"Labeling" refers to the covalent or noncovalent joining of a polynucleotide, polypeptide, or antibody with a reporter molecule capable of producing a detectable or measurable signal.

"Microarray" is any arrangement of nucleic acids, amino acids, antibodies, etc., on a substrate. The substrate may be a solid support such as beads, glass, paper, nitrocellulose, nylon, or an appropriate membrane.

"Linkers" are short stretches of nucleotide sequence which may be added to a vector or an mddt to create restriction endonuclease sites to facilitate cloning. "Polylinkers" are engineered to incorporate multiple restriction enzyme sites and to provide for the use of enzymes which leave 5' or 3' overhangs (e.g., BamHI, EcoRI, and HindIII) and those which provide blunt ends (e.g., EcoRV, SnaBI, and StuI).

"Naturally occurring" refers to an endogenous polynucleotide or polypeptide that may be isolated from viruses or prokaryotic or eukaryotic cells.

"Nucleic acid sequence" refers to the specific order of nucleotides joined by phosphodiester bonds in a linear, polymeric arrangement. Depending on the number of nucleotides, the nucleic acid sequence can be considered an oligomer, oligonucleotide, or polynucleotide. The nucleic acid can be DNA, RNA, or any nucleic acid analog, such as PNA, may be of genomic or synthetic origin, may be either double-stranded or single-stranded, and can represent either the sense or antisense (complementary) strand.

"Oligomer" refers to a nucleic acid sequence of at least about 6 nucleotides and as many as about 60 nucleotides, preferably about 15 to 40 nucleotides, and most preferably between about 20 and 30 nucleotides, that may be used in hybridization or amplification technologies. Oligomers may be used as, e.g., primers for PCR, and are usually chemically synthesized.

"Operably linked" refers to the situation in which a first nucleic acid sequence is placed in a functional relationship with the second nucleic acid sequence. For instance, a promoter is operably linked to a coding sequence if the promoter affects the transcription or expression of the coding sequence. Generally, operably linked DNA sequences may be in close proximity or contiguous and, where necessary to join two protein coding regions, in the same reading frame.

"Peptide nucleic acid" (PNA) refers to a DNA mimic in which nucleotide bases are attached to a pseudopeptide backbone to increase stability. PNAs, also designated antigene agents, can prevent gene expression by targeting complementary messenger RNA.

The phrases "percent identity" and "% identity", as applied to polynucleotide sequences, refer to the percentage of residue matches between at least two polynucleotide sequences aligned using a standardized algorithm. Such an algorithm may insert, in a standardized and reproducible way, gaps in the sequences being compared in order to optimize alignment between two sequences, and therefore achieve a more meaningful comparison of the two sequences.

Percent identity between polynucleotide sequences may be determined using the default parameters of the CLUSTAL V algorithm as incorporated into the MEGALIGN version 3.12e sequence alignment program. This program is part of the LASERGENE software package, a suite of molecular biological analysis programs (DNASTAR, Madison WI). CLUSTAL V is described in Higgins, D.G. and Sharp, P.M. (1989) CABIOS 5:151-153 and in Higgins, D.G. et al. (1992) CABIOS 8:189-191. For pairwise alignments of polynucleotide sequences, the default parameters are set as follows: Ktuple=2, gap penalty=5, window=4, and "diagonals saved"=4. The "weighted" residue weight table is selected as the default. Percent identity is reported by CLUSTAL V as the "percent similarity" between aligned polynucleotide sequence pairs.

Alternatively, a suite of commonly used and freely available sequence comparison algorithms is provided by the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) (Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410), which is available from several sources, including the NCBI, Bethesda, MD, and on the Internet at http://www.ncbi.nlm.nih.gov/BLAST/. The BLAST software suite includes various sequence analysis programs including "blastn," that is used to determine alignment between a known polynucleotide sequence and other sequences on a variety of databases. Also available is a tool called "BLAST 2 Sequences" that is used for direct pairwise comparison of two nucleotide sequences. "BLAST 2 Sequences" can be accessed and used interactively at http://www.ncbi.nlm.nih.gov/gorf/bl2/. The "BLAST 2 Sequences" tool can be used for both blastn and blastp (discussed below). BLAST programs are commonly used with gap and other parameters set to default settings. For example, to compare two nucleotide sequences, one may use blastn with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such default parameters may be, for example:

Matrix: BLOSUM62

Reward for match: 1

Penalty for mismatch: -2

Open Gap: 5 and Extension Gap: 2 penalties

Gap x drop-off: 50

Expect: 10

Word Size: 11

Filter: on

Percent identity may be measured over the length of an entire defined sequence, for example, as defined by a particular SEQ ID number, or may be measured over a shorter length, for example, over the length of a fragment taken from a larger, defined sequence, for instance, a fragment of at least 20, at least 30, at least 40, at least 50, at least 70, at least 100, or at least 200 contiguous nucleotides. Such lengths are exemplary only, and it is understood that any fragment length supported by the sequences shown herein, in figures or Sequence Listings, may be used to describe a length over which percentage identity may be measured.

Nucleic acid sequences that do not show a high degree of identity may nevertheless encode similar amino acid sequences due to the degeneracy of the genetic code. It is understood that changes in nucleic acid sequence can be made using this degeneracy to produce multiple nucleic acid sequences that all encode substantially the same protein.

The phrases "percent identity" and "% identity", as applied to polypeptide sequences, refer to the percentage of residue matches between at least two polypeptide sequences aligned using a standardized algorithm. Methods of polypeptide sequence alignment are well-known. Some alignment methods take into account conservative amino acid substitutions. Such conservative substitutions, explained in more detail above, generally preserve the hydrophobicity and acidity of the substituted residue, thus preserving the structure (and therefore function) of the folded polypeptide.

Percent identity between polypeptide sequences may be determined using the default parameters of the CLUSTAL V algorithm as incorporated into the MEGALIGN version 3.12e sequence alignment program (described and referenced above). For pairwise alignments of polypeptide sequences using CLUSTAL V, the default parameters are set as follows: Ktuple=1, gap penalty=3, window=5, and "diagonals saved"=5. The PAM250 matrix is selected as the default residue weight table. As with polynucleotide alignments, the percent identity is reported by CLUSTAL V as the "percent similarity" between aligned polypeptide sequence pairs.

Alternatively the NCBI BLAST software suite may be used. For example, for a pairwise comparison of two polypeptide sequences, one may use the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) with blastp set at default parameters. Such default parameters may be, for example:

Matrix: BLOSUM62

Open Gap: 11 and Extension Gap: 1 penalty

Gap x drop-off: 50

Expect: 10

Word Size: 3

Filter: on

Percent identity may be measured over the length of an entire defined polypeptide sequence, for example, as defined by a particular SEQ ID number, or may be measured over a shorter length, for example, over the length of a fragment taken from a larger, defined polypeptide sequence, for instance, a fragment of at least 15, at least 20, at least 30, at least 40, at least 50, at least 70 or at least 150 contiguous residues. Such lengths are exemplary only, and it is understood that any fragment length supported by the sequences shown herein, in figures or Sequence Listings, may be used to describe a length over which percentage identity may be measured.

"Post-translational modification" of an MDDT may involve lipidation, glycosylation, phosphorylation, acetylation, racemization, proteolytic cleavage, and other modifications known in the art. These processes may occur synthetically or biochemically. Biochemical modifications will vary by cell type depending on the enzymatic milieu and the MDDT.

"Probe" refers to mddt or fragments thereof, which are used to detect identical, allelic or related nucleic acid sequences. Probes are isolated oligonucleotides or polynucleotides attached to a detectable label or reporter molecule. Typical labels include radioactive isotopes, ligands, chemiluminescent agents, and enzymes. "Primers" are short nucleic acids, usually DNA oligonucleotides, which may be annealed to a target polynucleotide by complementary base-pairing. The primer may then be extended along the target DNA strand by a DNA polymerase enzyme. Primer pairs can be used for amplification (and identification) of a nucleic acid sequence, e.g., by the polymerase chain reaction (PCR).

Probes and primers as used in the present invention typically comprise at least 15 contiguous nucleotides of a known sequence. In order to enhance specificity, longer probes and primers may also be employed, such as probes and primers that comprise at least 20, 30, 40, 50, 60, 70, 80, 90, 100, or at least 150 consecutive nucleotides of the disclosed nucleic acid sequences. Probes and primers may be considerably longer than these examples, and it is understood that any length supported by the specification, including the figures and Sequence Listing, may be used.

Methods for preparing and using probes and primers are described in the references, for example Sambrook et al., 1989, Molecular Cloning: A Laboratory Manual, 2nd ed., vol. 1-3, Cold Spring Harbor Press, Plainview NY; Ausubel et al.,1987, Current Protocols in Molecular Biology, Greene Publ. Assoc. & Wiley-Intersciences, New York NY; Innis et al., 1990, PCR Protocols, A Guide to Methods and Applications, Academic Press, San Diego CA. PCR primer pairs can be derived from a known sequence, for example, by using computer programs intended for that purpose such as Primer (Version 0.5, 1991, Whitehead Institute for Biomedical Research, Cambridge MA).

Oligonucleotides for use as primers are selected using software known in the art for such purpose. For example, OLIGO 4.06 software is useful for the selection of PCR primer pairs of up to 100 nucleotides each, and for the analysis of oligonucleotides and larger polynucleotides of up to

5,000 nucleotides from an input polynucleotide sequence of up to 32 kilobases. Similar primer selection programs have incorporated additional features for expanded capabilities. For example, the PrimOU primer selection program (available to the public from the Genome Center at University of Texas South West Medical Center, Dallas TX) is capable of choosing specific primers from megabase sequences and is thus useful for designing primers on a genome-wide scope. The Primer3 primer selection program (available to the public from the Whitehead Institute/MIT Center for Genome Research, Cambridge MA) allows the user to input a "mispriming library," in which sequences to avoid as primer binding sites are user-specified. Primer3 is useful, in particular, for the selection of oligonucleotides for microarrays. (The source code for the latter two primer selection programs may also be obtained from their respective sources and modified to meet the user's specific needs.) The PrimeGen program (available to the public from the UK Human Genome Mapping Project Resource Centre, Cambridge UK) designs primers based on multiple sequence alignments, thereby allowing selection of primers that hybridize to either the most conserved or least conserved regions of aligned nucleic acid sequences. Hence, this program is useful for identification of both unique and conserved oligonucleotides and polynucleotide fragments. The oligonucleotides and polynucleotide fragments identified by any of the above selection methods are useful in hybridization technologies, for example, as PCR or sequencing primers, microarray elements, or specific probes to identify fully or partially complementary polynucleotides in a sample of nucleic acids. Methods of oligonucleotide selection are not limited to those described above.

"Purified" refers to molecules, either polynucleotides or polypeptides that are isolated or separated from their natural environment and are at least 60% free, preferably at least 75% free, and most preferably at least 90% free from other compounds with which they are naturally associated.

A "recombinant nucleic acid" is a sequence that is not naturally occurring or has a sequence that is made by an artificial combination of two or more otherwise separated segments of sequence. This artificial combination is often accomplished by chemical synthesis or, more commonly, by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques such as those described in Sambrook, supra. The term recombinant includes nucleic acids that have been altered solely by addition, substitution, or deletion of a portion of the nucleic acid. Frequently, a recombinant nucleic acid may include a nucleic acid sequence operably linked to a promoter sequence. Such a recombinant nucleic acid may be part of a vector that is used, for example, to transform a cell.

Alternatively, such recombinant nucleic acids may be part of a viral vector, e.g., based on a vaccinia virus, that could be use to vaccinate a mammal wherein the recombinant nucleic acid is expressed, inducing a protective immunological response in the mammal.

"Regulatory element" refers to a nucleic acid sequence from nontranslated regions of a gene, and includes enhancers, promoters, introns, and 3' untranslated regions, which interact with host proteins to carry out or regulate transcription or translation.

"Reporter" molecules are chemical or biochemical moieties used for labeling a nucleic acid, an amino acid, or an antibody. They include radionuclides; enzymes; fluorescent, chemiluminescent, or chromogenic agents; substrates; cofactors; inhibitors; magnetic particles; and other moieties known in the art.

An "RNA equivalent," in reference to a DNA sequence, is composed of the same linear sequence of nucleotides as the reference DNA sequence with the exception that all occurrences of the nitrogenous base thymine are replaced with uracil, and the sugar backbone is composed of ribose instead of deoxyribose.

"Sample" is used in its broadest sense. Samples may contain nucleic or amino acids, antibodies, or other materials, and may be derived from any source (e.g., bodily fluids including, but not limited to, saliva, blood, and urine; chromosome(s), organelles, or membranes isolated from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; and cleared cells or tissues or blots or imprints from such cells or tissues).

"Specific binding" or "specifically binding" refers to the interaction between a protein or peptide and its agonist, antibody, antagonist, or other binding partner. The interaction is dependent upon the presence of a particular structure of the protein, e.g., the antigenic determinant or epitope, recognized by the binding molecule. For example, if an antibody is specific for epitope "A," the presence of a polypeptide containing epitope A, or the presence of free unlabeled A, in a reaction containing free labeled A and the antibody will reduce the amount of labeled A that binds to the antibody.

"Substitution" refers to the replacement of at least one nucleotide or amino acid by a different nucleotide or amino acid.

"Substrate" refers to any suitable rigid or semi-rigid support including, e.g., membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, tubing, plates, polymers, microparticles or capillaries. The substrate can have a variety of surface forms, such as wells, trenches, pins, channels and pores, to which polynucleotides or polypeptides are bound.

A "transcript image" refers to the collective pattern of gene expression by a particular tissue or cell type under given conditions at a given time.

"Transformation" refers to a process by which exogenous DNA enters a recipient cell. Transformation may occur under natural or artificial conditions using various methods well known in the art. Transformation may rely on any known method for the insertion of foreign nucleic acid

sequences into a prokaryotic or eukaryotic host cell. The method is selected based on the host cell being transformed.

"Transformants" include stably transformed cells in which the inserted DNA is capable of replication either as an autonomously replicating plasmid or as part of the host chromosome, as well as cells which transiently express inserted DNA or RNA.

A "transgenic organism," as used herein, is any organism, including but not limited to animals and plants, in which one or more of the cells of the organism contains heterologous nucleic acid introduced by way of human intervention, such as by transgenic techniques well known in the art. The nucleic acid is introduced into the cell, directly or indirectly by introduction into a precursor of the cell, by way of deliberate genetic manipulation, such as by microinjection or by infection with a recombinant virus. The term genetic manipulation does not include classical cross-breeding, or in vitro fertilization, but rather is directed to the introduction of a recombinant DNA molecule. The transgenic organisms contemplated in accordance with the present invention include bacteria, cyanobacteria, fungi, and plants and animals. The isolated DNA of the present invention can be introduced into the host by methods known in the art, for example infection, transfection, transformation or transconjugation. Techniques for transferring the DNA of the present invention into such organisms are widely known and provided in references such as Sambrook et al. (1989), supra.

A "variant" of a particular nucleic acid sequence is defined as a nucleic acid sequence having at least 25% sequence identity to the particular nucleic acid sequence over a certain length of one of the nucleic acid sequences using blastn with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such a pair of nucleic acids may show, for example, at least 30%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95% or even at least 98% or greater sequence identity over a certain defined length. The variant may result in "conservative" amino acid changes which do not affect structural and/or chemical properties. A variant may be described as, for example, an "allelic" (as defined above), "splice," "species," or "polymorphic" variant. A splice variant may have significant identity to a reference molecule, but will generally have a greater or lesser number of polynucleotides due to alternate splicing of exons during mRNA processing. The corresponding polypeptide may possess additional functional domains or lack domains that are present in the reference molecule. Species variants are polynucleotide sequences that vary from one species to another. The resulting polypeptides generally will have significant amino acid identity relative to each other. A polymorphic variant is a variation in the polynucleotide sequence of a particular gene between individuals of a given species. Polymorphic variants also may encompass "single nucleotide polymorphisms" (SNPs) in which the polynucleotide sequence varies by one base. The presence of SNPs may be indicative of, for example, a certain population, a disease

state, or a propensity for a disease state.

In an alternative, variants of the polynucleotides of the present invention may be generated through recombinant methods. One possible method is a DNA shuffling technique such as MOLECULARBREEDING (Maxygen Inc., Santa Clara CA; described in U.S. Patent Number 5,837,458; Chang, C.-C. et al. (1999) Nat. Biotechnol. 17:793-797; Christians, F.C. et al. (1999) Nat. Biotechnol. 17:259-264; and Crameri, A. et al. (1996) Nat. Biotechnol. 14:315-319) to alter or improve the biological properties of MDDT, such as its biological or enzymatic activity or its ability to bind to other molecules or compounds. DNA shuffling is a process by which a library of gene variants is produced using PCR-mediated recombination of gene fragments. The library is then subjected to selection or screening procedures that identify those gene variants with the desired properties. These preferred variants may then be pooled and further subjected to recursive rounds of DNA shuffling and selection/screening. Thus, genetic diversity is created through "artificial" breeding and rapid molecular evolution. For example, fragments of a single gene containing random point mutations may be recombined, screened, and then reshuffled until the desired properties are optimized. Alternatively, fragments of a given gene may be recombined with fragments of homologous genes in the same gene family, either from the same or different species, thereby maximizing the genetic diversity of multiple naturally occurring genes in a directed and controllable manner.

A "variant" of a particular polypeptide sequence is defined as a polypeptide sequence having at least 40% sequence identity to the particular polypeptide sequence over a certain length of one of the polypeptide sequences using blastp with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such a pair of polypeptides may show, for example, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%, or at least 98% or greater sequence identity over a certain defined length of one of the polypeptides.

THE INVENTION

In a particular embodiment, cDNA sequences derived from human tissues and cell lines were aligned based on nucleotide sequence identity and assembled into "consensus" or "template" sequences which are designated by the template identification numbers (template IDs) in column 2 of Table 1. The sequence identification numbers (SEQ ID NO:s) corresponding to the template IDs are shown in column 1. The template sequences have similarity to GenBank sequences, or "hits," as designated by the GI Numbers in column 3. The statistical probability of each GenBank hit is indicated by a probability score in column 4, and the functional annotation corresponding to each GenBank hit is listed in column 5.

The invention incorporates the nucleic acid sequences of these templates as disclosed in the

Sequence Listing and the use of these sequences in the diagnosis and treatment of disease states characterized by defects in molecules for disease detection and treatment. The invention further utilizes these sequences in hybridization and amplification technologies, and in particular, in technologies which assess gene expression patterns correlated with specific cells or tissues and their responses in vivo or in vitro to pharmaceutical agents, toxins, and other treatments. In this manner, the sequences of the present invention are used to develop a transcript image for a particular cell or tissue.

Derivation of Nucleic Acid Sequences

cDNA was isolated from libraries constructed using RNA derived from normal and diseased human tissues and cell lines. The human tissues and cell lines used for cDNA library construction were selected from a broad range of sources to provide a diverse population of cDNAs representative of gene transcription throughout the human body. Descriptions of the human tissues and cell lines used for cDNA library construction are provided in the LIFESEQ database (Incyte Genomics, Inc. (Incyte), Palo Alto CA). Human tissues were broadly selected from, for example, cardiovascular, dermatologic, endocrine, gastrointestinal, hematopoietic/immune system, musculoskeletal, neural, reproductive, and urologic sources.

Cell lines used for cDNA library construction were derived from, for example, leukemic cells, teratocarcinomas, neuroepitheliomas, cervical carcinoma, lung fibroblasts, and endothelial cells. Such cell lines include, for example, THP-1, Jurkat, HUVEC, hNT2, WI38, HeLa, and other cell lines commonly used and available from public depositories (American Type Culture Collection, Manassas VA). Prior to mRNA isolation, cell lines were untreated, treated with a pharmaceutical agent such as 5'-aza-2'-deoxycytidine, treated with an activating agent such as lipopolysaccharide in the case of leukocytic cell lines, or, in the case of endothelial cell lines, subjected to shear stress.

Sequencing of the cDNAs

Methods for DNA sequencing are well known in the art. Conventional enzymatic methods employ the Klenow fragment of DNA polymerase I, SEQUENASE DNA polymerase (U.S. Biochemical Corporation, Cleveland OH), Taq polymerase (PE Biosystems, Foster City CA), thermostable T7 polymerase (Amersham Pharmacia Biotech, Inc. (Amersham Pharmacia Biotech), Piscataway NJ), or combinations of polymerases and proofreading exonucleases such as those found in the ELONGASE amplification system (Life Technologies Inc. (Life Technologies), Gaithersburg MD), to extend the nucleic acid sequence from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single-stranded and double-stranded templates. Chain termination reaction products may be electrophoresed on urea-

polyacrylamide gels and detected either by autoradiography (for radioisotope-labeled nucleotides) or by fluorescence (for fluorophore-labeled nucleotides). Automated methods for mechanized reaction preparation, sequencing, and analysis using fluorescence detection methods have been developed. Machines used to prepare cDNAs for sequencing can include the MICROLAB 2200 liquid transfer system (Hamilton Company (Hamilton), Reno NV), Peltier thermal cycler (PTC200; MJ Research, Inc. (MJ Research), Watertown MA), and ABI CATALYST 800 thermal cycler (PE Biosystems). Sequencing can be carried out using, for example, the ABI 373 or 377 (PE Biosystems) or MEGABACE 1000 (Molecular Dynamics, Inc. (Molecular Dynamics), Sunnyvale CA) DNA sequencing systems, or other automated and manual sequencing systems well known in the art.

The nucleotide sequences of the Sequence Listing have been prepared by current, state-of-the-art, automated methods and, as such, may contain occasional sequencing errors or unidentified nucleotides. Such unidentified nucleotides are designated by an N. These infrequent unidentified bases do not represent a hindrance to practicing the invention for those skilled in the art. Several methods employing standard recombinant techniques may be used to correct errors and complete the missing sequence information. (See, e.g., those described in Ausubel, F.M. et al. (1997) Short Protocols in Molecular Biology, John Wiley & Sons, New York NY; and Sambrook, J. et al. (1989) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview NY.)

Assembly of cDNA Sequences

Human polynucleotide sequences may be assembled using programs or algorithms well known in the art. Sequences to be assembled are related, wholly or in part, and may be derived from a single or many different transcripts. Assembly of the sequences can be performed using such programs as PHRAP (Phils Revised Assembly Program) and the GELVIEW fragment assembly system (GCG), or other methods known in the art.

Alternatively, cDNA sequences are used as "component" sequences that are assembled into "template" or "consensus" sequences as follows. Sequence chromatograms are processed, verified, and quality scores are obtained using PHRED. Raw sequences are edited using an editing pathway known as Block 1 (See, e.g., the LIFESEQ Assembled User Guide, Incyte Genomics, Palo Alto, CA). A series of BLAST comparisons is performed and low-information segments and repetitive elements (e.g., dinucleotide repeats, Alu repeats, etc.) are replaced by "n's", or masked, to prevent spurious matches. Mitochondrial and ribosomal RNA sequences are also removed. The processed sequences are then loaded into a relational database management system (RDMS) which assigns edited sequences to existing templates, if available. When additional sequences are added into the RDMS, a process is initiated which modifies existing templates or creates new templates from works in progress (i.e., nonfinal assembled sequences) containing queued sequences or the sequences

themselves. After the new sequences have been assigned to templates, the templates can be merged into bins. If multiple templates exist in one bin, the bin can be split and the templates reannotated.

Once gene bins have been generated based upon sequence alignments, bins are "clone joined" based upon clone information. Clone joining occurs when the 5' sequence of one clone is present in one bin and the 3' sequence from the same clone is present in a different bin, indicating that the two bins should be merged into a single bin. Only bins which share at least two different clones are merged.

A resultant template sequence may contain either a partial or a full length open reading frame, or all or part of a genetic regulatory element. This variation is due in part to the fact that the full length cDNAs of many genes are several hundred, and sometimes several thousand, bases in length. With current technology, cDNAs comprising the coding regions of large genes cannot be cloned because of vector limitations, incomplete reverse transcription of the mRNA, or incomplete "second strand" synthesis. Template sequences may be extended to include additional contiguous sequences derived from the parent RNA transcript using a variety of methods known to those of skill in the art. Extension may thus be used to achieve the full length coding sequence of a gene.

Analysis of the cDNA Sequences

The cDNA sequences are analyzed using a variety of programs and algorithms which are well known in the art. (See, e.g., Ausubel, 1997, supra, Chapter 7.7; Meyers, R.A. (Ed.) (1995) Molecular Biology and Biotechnology, Wiley VCH, New York NY, pp. 856-853; and Table 5.) These analyses comprise both reading frame determinations, e.g., based on triplet codon periodicity for particular organisms (Fickett, J.W. (1982) Nucleic Acids Res. 10:5303-5318); analyses of potential start and stop codons; and homology searches.

Computer programs known to those of skill in the art for performing computer-assisted searches for amino acid and nucleic acid sequence similarity, include, for example, Basic Local Alignment Search Tool (BLAST; Altschul, S.F. (1993) J. Mol. Evol. 36:290-300; Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410). BLAST is especially useful in determining exact matches and comparing two sequence fragments of arbitrary but equal lengths, whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score set by the user (Karlin, S. et al. (1988) Proc. Natl. Acad. Sci. USA 85:841-845). Using an appropriate search tool (e.g., BLAST or HMM), GenBank, SwissProt, BLOCKS, PFAM and other databases may be searched for sequences containing regions of homology to a query mddt or MDDT of the present invention.

Other approaches to the identification, assembly, storage, and display of nucleotide and polypeptide sequences are provided in "Relational Database for Storing Biomolecule Information,"

U.S.S.N. 08/947,845, filed October 9, 1997; "Project-Based Full-Length Biomolecular Sequence Database," U.S.S.N. 08/811,758, filed March 6, 1997; and "Relational Database and System for Storing Information Relating to Biomolecular Sequences," U.S.S.N. 09/034,807, filed March 4, 1998, all of which are incorporated by reference herein in their entirety.

Protein hierarchies can be assigned to the putative encoded polypeptide based on, e.g., motif, BLAST, or biological analysis. Methods for assigning these hierarchies are described, for example, in "Database System Employing Protein Function Hierarchies for Viewing Biomolecular Sequence Data," U.S.S.N. 08/812,290, filed March 6, 1997, incorporated herein by reference.

Human Disease Detection and Treatment Molecule Sequences

The mddt of the present invention may be used for a variety of diagnostic and therapeutic purposes. For example, an mddt may be used to diagnose a particular condition, disease, or disorder associated with disease detection and treatment molecules. Such conditions, diseases, and disorders include, but are not limited to, a cell proliferative disorder, such as actinic keratosis, arteriosclerosis, atherosclerosis, bursitis, cirrhosis, hepatitis, mixed connective tissue disease (MCTD), myelofibrosis, paroxysmal nocturnal hemoglobinuria, polycythemia vera, psoriasis, primary thrombocythemia, and cancers including adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, a cancer of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus; and an autoimmune/inflammatory disorder, such as actinic keratosis, acquired immunodeficiency syndrome (AIDS), Addison's disease, adult respiratory distress syndrome, allergies, ankylosing spondylitis, amyloidosis, anemia, arteriosclerosis, asthma, atherosclerosis, autoimmune hemolytic anemia, autoimmune thyroiditis, bronchitis, bursitis, cholecystitis, cirrhosis, contact dermatitis, Crohn's disease, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, erythroblastosis fetalis, erythema nodosum, atrophic gastritis, glomerulonephritis, Goodpasture's syndrome, gout, Graves' disease, Hashimoto's thyroiditis, paroxysmal nocturnal hemoglobinuria, hepatitis, hypereosinophilia, irritable bowel syndrome, episodic lymphopenia with lymphocytotoxins, mixed connective tissue disease (MCTD), multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, myelofibrosis, osteoarthritis, osteoporosis, pancreatitis, polycythemia vera, polymyositis, psoriasis, Reiter's syndrome, rheumatoid arthritis, scleroderma, Sjögren's syndrome, systemic anaphylaxis, systemic lupus erythematosus, systemic sclerosis, primary thrombocythemia, thrombocytopenic purpura, ulcerative colitis, uveitis, Werner syndrome, complications of cancer, hemodialysis, and extracorporeal circulation, trauma, and hematopoietic cancer including lymphoma, leukemia, and myeloma. The mddt can be used to detect the presence

of, or to quantify the amount of, an mddt-related polynucleotide in a sample. This information is then compared to information obtained from appropriate reference samples, and a diagnosis is established. Alternatively, a polynucleotide complementary to a given mddt can inhibit or inactivate a therapeutically relevant gene related to the mddt.

Analysis of mddt Expression Patterns

The expression of mddt may be routinely assessed by hybridization-based methods to determine, for example, the tissue-specificity, disease-specificity, or developmental stage-specificity of mddt expression. For example, the level of expression of mddt may be compared among different cell types or tissues, among diseased and normal cell types or tissues, among cell types or tissues at different developmental stages, or among cell types or tissues undergoing various treatments. This type of analysis is useful, for example, to assess the relative levels of mddt expression in fully or partially differentiated cells or tissues, to determine if changes in mddt expression levels are correlated with the development or progression of specific disease states, and to assess the response of a cell or tissue to a specific therapy, for example, in pharmacological or toxicological studies. Methods for the analysis of mddt expression are based on hybridization and amplification technologies and include membrane-based procedures such as northern blot analysis, high-throughput procedures that utilize, for example, microarrays, and PCR-based procedures.

Hybridization and Genetic Analysis

The mddt, their fragments, or complementary sequences, may be used to identify the presence of and/or to determine the degree of similarity between two (or more) nucleic acid sequences. The mddt may be hybridized to naturally occurring or recombinant nucleic acid sequences under appropriately selected temperatures and salt concentrations. Hybridization with a probe based on the nucleic acid sequence of at least one of the mddt allows for the detection of nucleic acid sequences, including genomic sequences, which are identical or related to the mddt of the Sequence Listing. Probes may be selected from non-conserved or unique regions of at least one of the polynucleotides of SEQ ID NO:1-14 and tested for their ability to identify or amplify the target nucleic acid sequence using standard protocols.

Polynucleotide sequences that are capable of hybridizing, in particular, to those shown in SEQ ID NO:1-14 and fragments thereof, can be identified using various conditions of stringency. (See, e.g., Wahl, G.M. and S.L. Berger (1987) Methods Enzymol. 152:399-407; Kimmel, A.R. (1987) Methods Enzymol. 152:507-511.) Hybridization conditions are discussed in "Definitions."

A probe for use in Southern or northern hybridization may be derived from a fragment of an mddt sequence, or its complement, that is up to several hundred nucleotides in length and is either

single-stranded or double-stranded. Such probes may be hybridized in solution to biological materials such as plasmids, bacterial, yeast, or human artificial chromosomes, cleared or sectioned tissues, or to artificial substrates containing mddt. Microarrays are particularly suitable for identifying the presence of and detecting the level of expression for multiple genes of interest by examining gene expression correlated with, e.g., various stages of development, treatment with a drug or compound, or disease progression. An array analogous to a dot or slot blot may be used to arrange and link polynucleotides to the surface of a substrate using one or more of the following: mechanical (vacuum), chemical, thermal, or UV bonding procedures. Such an array may contain any number of mddt and may be produced by hand or by using available devices, materials, and machines.

Microarrays may be prepared, used, and analyzed using methods known in the art. (See, e.g., Brennan, T.M. et al. (1995) U.S. Patent No. 5,474,796; Schena, M. et al. (1996) Proc. Natl. Acad. Sci. USA 93:10614-10619; Baldeschweiler et al. (1995) PCT application WO95/251116; Shalon, D. et al. (1995) PCT application WO95/35505; Heller, R.A. et al. (1997) Proc. Natl. Acad. Sci. USA 94:2150-2155; and Heller, M.J. et al. (1997) U.S. Patent No. 5,605,662.)

Probes may be labeled by either PCR or enzymatic techniques using a variety of commercially available reporter molecules. For example, commercial kits are available for radioactive and chemiluminescent labeling (Amersham Pharmacia Biotech) and for alkaline phosphatase labeling (Life Technologies). Alternatively, mddt may be cloned into commercially available vectors for the production of RNA probes. Such probes may be transcribed in the presence of at least one labeled nucleotide (e.g., ^{32}P-ATP, Amersham Pharmacia Biotech).

Additionally the polynucleotides of SEQ ID NO:1-14 or suitable fragments thereof can be used to isolate full length cDNA sequences utilizing hybridization and/or amplification procedures well known in the art, e.g., cDNA library screening, PCR amplification, etc. The molecular cloning of such full length cDNA sequences may employ the method of cDNA library screening with probes using the hybridization, stringency, washing, and probing strategies described above and in Ausubel, supra, Chapters 3, 5, and 6. These procedures may also be employed with genomic libraries to isolate genomic sequences of mddt in order to analyze, e.g., regulatory elements.

Genetic Mapping

Gene identification and mapping are important in the investigation and treatment of almost all conditions, diseases, and disorders. Cancer, cardiovascular disease, Alzheimer's disease, arthritis, diabetes, and mental illnesses are of particular interest. Each of these conditions is more complex than the single gene defects of sickle cell anemia or cystic fibrosis, with select groups of genes being predictive of predisposition for a particular condition, disease, or disorder. For example, cardiovascular disease may result from malfunctioning receptor molecules that fail to clear

cholesterol from the bloodstream, and diabetes may result when a particular individual's immune system is activated by an infection and attacks the insulin-producing cells of the pancreas. In some studies, Alzheimer's disease has been linked to a gene on chromosome 21; other studies predict a different gene and location. Mapping of disease genes is a complex and reiterative process and generally proceeds from genetic linkage analysis to physical mapping.

As a condition is noted among members of a family, a genetic linkage map traces parts of chromosomes that are inherited in the same pattern as the condition. Statistics link the inheritance of particular conditions to particular regions of chromosomes, as defined by RFLP or other markers. (See, for example, Lander, E. S. and Botstein, D. (1986) Proc. Natl. Acad. Sci. USA 83:7353-7357.) Occasionally, genetic markers and their locations are known from previous studies. More often, however, the markers are simply stretches of DNA that differ among individuals. Examples of genetic linkage maps can be found in various scientific journals or at the Online Mendelian Inheritance in Man (OMIM) World Wide Web site.

In another embodiment of the invention, mddt sequences may be used to generate hybridization probes useful in chromosomal mapping of naturally occurring genomic sequences. Either coding or noncoding sequences of mddt may be used, and in some instances, noncoding sequences may be preferable over coding sequences. For example, conservation of an mddt coding sequence among members of a multi-gene family may potentially cause undesired cross hybridization during chromosomal mapping. The sequences may be mapped to a particular chromosome, to a specific region of a chromosome, or to artificial chromosome constructions, e.g., human artificial chromosomes (HACs), yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), bacterial P1 constructions, or single chromosome cDNA libraries. (See, e.g., Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355; Price, C.M. (1993) Blood Rev. 7:127-134; and Trask, B.J. (1991) Trends Genet. 7:149-154.)

Fluorescent in situ hybridization (FISH) may be correlated with other physical chromosome mapping techniques and genetic map data. (See, e.g., Meyers, supra, pp. 965-968.) Correlation between the location of mddt on a physical chromosomal map and a specific disorder, or a predisposition to a specific disorder, may help define the region of DNA associated with that disorder. The mddt sequences may also be used to detect polymorphisms that are genetically linked to the inheritance of a particular condition, disease, or disorder.

In situ hybridization of chromosomal preparations and genetic mapping techniques, such as linkage analysis using established chromosomal markers, may be used for extending existing genetic maps. Often the placement of a gene on the chromosome of another mammalian species, such as mouse, may reveal associated markers even if the number or arm of the corresponding human chromosome is not known. These new marker sequences can be mapped to human chromosomes and

may provide valuable information to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once a disease or syndrome has been crudely correlated by genetic linkage with a particular genomic region, e.g., ataxia-telangiectasia to 11q22-23, any sequences mapping to that area may represent associated or regulatory genes for further investigation. (See, e.g., Gatti, R.A. et al. (1988) Nature 336:577-580.) The nucleotide sequences of the subject invention may also be used to detect differences in chromosomal architecture due to translocation, inversion, etc., among normal, carrier, or affected individuals.

Once a disease-associated gene is mapped to a chromosomal region, the gene must be cloned in order to identify mutations or other alterations (e.g., translocations or inversions) that may be correlated with disease. This process requires a physical map of the chromosomal region containing the disease-gene of interest along with associated markers. A physical map is necessary for determining the nucleotide sequence of and order of marker genes on a particular chromosomal region. Physical mapping techniques are well known in the art and require the generation of overlapping sets of cloned DNA fragments from a particular organelle, chromosome, or genome. These clones are analyzed to reconstruct and catalog their order. Once the position of a marker is determined, the DNA from that region is obtained by consulting the catalog and selecting clones from that region. The gene of interest is located through positional cloning techniques using hybridization or similar methods.

Diagnostic Uses

The mddt of the present invention may be used to design probes useful in diagnostic assays. Such assays, well known to those skilled in the art, may be used to detect or confirm conditions, disorders, or diseases associated with abnormal levels of mddt expression. Labeled probes developed from mddt sequences are added to a sample under hybridizing conditions of desired stringency. In some instances, mddt, or fragments or oligonucleotides derived from mddt, may be used as primers in amplification steps prior to hybridization. The amount of hybridization complex formed is quantified and compared with standards for that cell or tissue. If mddt expression varies significantly from the standard, the assay indicates the presence of the condition, disorder, or disease. Qualitative or quantitative diagnostic methods may include northern, dot blot, or other membrane or dip-stick based technologies or multiple-sample format technologies such as PCR, enzyme-linked immunosorbent assay (ELISA)-like, pin, or chip-based assays.

The probes described above may also be used to monitor the progress of conditions, disorders, or diseases associated with abnormal levels of mddt expression, or to evaluate the efficacy of a particular therapeutic treatment. The candidate probe may be identified from the mddt that are specific to a given human tissue and have not been observed in GenBank or other genome databases.

Such a probe may be used in animal studies, preclinical tests, clinical trials, or in monitoring the treatment of an individual patient. In a typical process, standard expression is established by methods well known in the art for use as a basis of comparison, samples from patients affected by the disorder or disease are combined with the probe to evaluate any deviation from the standard profile, and a therapeutic agent is administered and effects are monitored to generate a treatment profile. Efficacy is evaluated by determining whether the expression progresses toward or returns to the standard normal pattern. Treatment profiles may be generated over a period of several days or several months. Statistical methods well known to those skilled in the art may be use to determine the significance of such therapeutic agents.

The polynucleotides are also useful for identifying individuals from minute biological samples, for example, by matching the RFLP pattern of a sample's DNA to that of an individual's DNA. The polynucleotides of the present invention can also be used to determine the actual base-by-base DNA sequence of selected portions of an individual's genome. These sequences can be used to prepare PCR primers for amplifying and isolating such selected DNA, which can then be sequenced. Using this technique, an individual can be identified through a unique set of DNA sequences. Once a unique ID database is established for an individual, positive identification of that individual can be made from extremely small tissue samples.

In a particular aspect, oligonucleotide primers derived from the mddt of the invention may be used to detect single nucleotide polymorphisms (SNPs). SNPs are substitutions, insertions and deletions that are a frequent cause of inherited or acquired genetic disease in humans. Methods of SNP detection include, but are not limited to, single-stranded conformation polymorphism (SSCP) and fluorescent SSCP (fSSCP) methods. In SSCP, oligonucleotide primers derived from the polynucleotide sequences encoding MDDT are used to amplify DNA using the polymerase chain reaction (PCR). The DNA may be derived, for example, from diseased or normal tissue, biopsy samples, bodily fluids, and the like. SNPs in the DNA cause differences in the secondary and tertiary structures of PCR products in single-stranded form, and these differences are detectable using gel electrophoresis in non-denaturing gels. In fSCCP, the oligonucleotide primers are fluorescently labeled, which allows detection of the amplimers in high-throughput equipment such as DNA sequencing machines. Additionally, sequence database analysis methods, termed in silico SNP (isSNP), are capable of identifying polymorphisms by comparing the sequence of individual overlapping DNA fragments which assemble into common consensus sequences. These computer-based methods filter out sequence variations due to laboratory preparation of DNA and sequencing errors using statistical models and automated analyses of DNA sequence chromatograms. In the alternative, SNPs may be detected and characterized by mass spectrometry using, for example, the high throughput MASSARRAY system (Sequenom, Inc., San Diego CA).

DNA-based identification techniques are critical in forensic technology. DNA sequences taken from very small biological samples such as tissues, e.g., hair or skin, or body fluids, e.g., blood, saliva, semen, etc., can be amplified using, e.g., PCR, to identify individuals. (See, e.g., Erlich, H. (1992) PCR Technology, Freeman and Co., New York, NY). Similarly, polynucleotides of the present invention can be used as polymorphic markers.

There is also a need for reagents capable of identifying the source of a particular tissue. Appropriate reagents can comprise, for example, DNA probes or primers prepared from the sequences of the present invention that are specific for particular tissues. Panels of such reagents can identify tissue by species and/or by organ type. In a similar fashion, these reagents can be used to screen tissue cultures for contamination.

The polynucleotides of the present invention can also be used as molecular weight markers on nucleic acid gels or Southern blots, as diagnostic probes for the presence of a specific mRNA in a particular cell type, in the creation of subtracted cDNA libraries which aid in the discovery of novel polynucleotides, in selection and synthesis of oligomers for attachment to an array or other support, and as an antigen to elicit an immune response.

Disease Model Systems Using mddt

The mddt of the invention or their mammalian homologs may be "knocked out" in an animal model system using homologous recombination in embryonic stem (ES) cells. Such techniques are well known in the art and are useful for the generation of animal models of human disease. (See, e.g., U.S. Patent Number 5,175,383 and U.S. Patent Number 5,767,337.) For example, mouse ES cells, such as the mouse 129/SvJ cell line, are derived from the early mouse embryo and grown in culture. The ES cells are transformed with a vector containing the gene of interest disrupted by a marker gene, e.g., the neomycin phosphotransferase gene (neo; Capecchi, M.R. (1989) Science 244:1288-1292). The vector integrates into the corresponding region of the host genome by homologous recombination. Alternatively, homologous recombination takes place using the Cre-loxP system to knockout a gene of interest in a tissue- or developmental stage-specific manner (Marth, J.D. (1996) Clin. Invest. 97:1999-2002; Wagner, K.U. et al. (1997) Nucleic Acids Res. 25:4323-4330). Transformed ES cells are identified and microinjected into mouse cell blastocysts such as those from the C57BL/6 mouse strain. The blastocysts are surgically transferred to pseudopregnant dams, and the resulting chimeric progeny are genotyped and bred to produce heterozygous or homozygous strains. Transgenic animals thus generated may be tested with potential therapeutic or toxic agents.

The mddt of the invention may also be manipulated in vitro in ES cells derived from human blastocysts. Human ES cells have the potential to differentiate into at least eight separate cell lineages including endoderm, mesoderm, and ectodermal cell types. These cell lineages differentiate

into, for example, neural cells, hematopoietic lineages, and cardiomyocytes (Thomson, J.A. et al. (1998) Science 282:1145-1147).

The mddt of the invention can also be used to create "knockin" humanized animals (pigs) or transgenic animals (mice or rats) to model human disease. With knockin technology, a region of mddt is injected into animal ES cells, and the injected sequence integrates into the animal cell genome. Transformed cells are injected into blastulae, and the blastulae are implanted as described above. Transgenic progeny or inbred lines are studied and treated with potential pharmaceutical agents to obtain information on treatment of a human disease. Alternatively, a mammal inbred to overexpress mddt, resulting, e.g., in the secretion of MDDT in its milk, may also serve as a convenient source of that protein (Janne, J. et al. (1998) Biotechnol. Annu. Rev. 4:55-74).

Screening Assays

MDDT encoded by polynucleotides of the present invention may be used to screen for molecules that bind to or are bound by the encoded polypeptides. The binding of the polypeptide and the molecule may activate (agonist), increase, inhibit (antagonist), or decrease activity of the polypeptide or the bound molecule. Examples of such molecules include antibodies, oligonucleotides, proteins (e.g., receptors), or small molecules.

Preferably, the molecule is closely related to the natural ligand of the polypeptide, e.g., a ligand or fragment thereof, a natural substrate, or a structural or functional mimetic. (See, Coligan et al., (1991) Current Protocols in Immunology 1(2): Chapter 5.) Similarly, the molecule can be closely related to the natural receptor to which the polypeptide binds, or to at least a fragment of the receptor, e.g., the active site. In either case, the molecule can be rationally designed using known techniques. Preferably, the screening for these molecules involves producing appropriate cells which express the polypeptide, either as a secreted protein or on the cell membrane. Preferred cells include cells from mammals, yeast, Drosophila, or E. coli. Cells expressing the polypeptide or cell membrane fractions which contain the expressed polypeptide are then contacted with a test compound and binding, stimulation, or inhibition of activity of either the polypeptide or the molecule is analyzed.

An assay may simply test binding of a candidate compound to the polypeptide, wherein binding is detected by a fluorophore, radioisotope, enzyme conjugate, or other detectable label. Alternatively, the assay may assess binding in the presence of a labeled competitor.

Additionally, the assay can be carried out using cell-free preparations, polypeptide/molecule affixed to a solid support, chemical libraries, or natural product mixtures. The assay may also simply comprise the steps of mixing a candidate compound with a solution containing a polypeptide, measuring polypeptide/molecule activity or binding, and comparing the polypeptide/molecule activity or binding to a standard.

Preferably, an ELISA assay using, e.g., a monoclonal or polyclonal antibody, can measure polypeptide level in a sample. The antibody can measure polypeptide level by either binding, directly or indirectly, to the polypeptide or by competing with the polypeptide for a substrate.

All of the above assays can be used in a diagnostic or prognostic context. The molecules discovered using these assays can be used to treat disease or to bring about a particular result in a patient (e.g., blood vessel growth) by activating or inhibiting the polypeptide/molecule. Moreover, the assays can discover agents which may inhibit or enhance the production of the polypeptide from suitably manipulated cells or tissues.

Transcript Imaging

Another embodiment relates to the use of mddt to develop a transcript image of a tissue or cell type. A transcript image is the collective pattern of gene expression by a particular tissue or cell type under given conditions and at a given time. This pattern of gene expression is defined by the number of expressed genes, their abundance, and their function. Thus the mddt of the present invention may be used to develop a transcript image of a tissue or cell type by hybridizing, preferably in a microarray format, the mddt of the present invention to the totality of transcripts or reverse transcripts of a tissue or cell type. The resultant transcript image would provide a profile of gene activity pertaining to disease detection and treatment.

Transcript images which profile mddt expression may be generated using transcripts isolated from tissues, cell lines, biopsies, or other biological samples. The transcript image may thus reflect mddt expression in vivo, as in the case of a tissue or biopsy sample, or in vitro, as in the case of a cell line. Transcript images may be used to profile mddt expression in distinct tissue types. This process can be used to determine disease detection and treatment molecule activity in a particular tissue type relative to this activity in a different tissue type. Transcript images may be used to generate a profile of mddt expression characteristic of diseased tissue. Transcript images of tissues before and after treatment may be used for diagnostic purposes, to monitor the progression of disease, and to monitor the efficacy of drug treatments for diseases which affect the activity of disease detection and treatment molecules.

Transcript images which profile mddt expression may also be used in conjunction with in vitro model systems and preclinical evaluation of pharmaceuticals. Transcript images of cell lines can be used to assess disease detection and treatment molecule activity and/or to identify cell lines that lack or misregulate this activity. Such cell lines may then be treated with pharmaceutical agents, and a transcript image following treatment may indicate the efficacy of these agents in restoring desired levels of this activity. A similar approach may be used to assess the toxicity of pharmaceutical agents as reflected by undesirable changes in disease detection and treatment

molecule activity. Candidate pharmaceutical agents may be evaluated by comparing their associated transcript images with those of pharmaceutical agents of known effectiveness.

Antisense Molecules

The polynucleotides of the present invention are useful in antisense technology. Antisense technology or therapy relies on the modulation of expression of a target protein through the specific binding of an antisense sequence to a target sequence encoding the target protein or directing its expression. (See, e.g., Agrawal, S., ed. (1996) Antisense Therapeutics, Humana Press Inc., Totawa NJ; Alama, A. et al. (1997) Pharmacol. Res. 36(3):171-178; Crooke, S.T. (1997) Adv. Pharmacol. 40:1-49; Sharma, H.W. and R. Narayanan (1995) Bioessays 17(12):1055-1063; and Lavrosky, Y. et al. (1997) Biochem. Mol. Med. 62(1):11-22.) An antisense sequence is a polynucleotide sequence capable of specifically hybridizing to at least a portion of the target sequence. Antisense sequences bind to cellular mRNA and/or genomic DNA, affecting translation and/or transcription. Antisense sequences can be DNA, RNA, or nucleic acid mimics and analogs. (See, e.g., Rossi, J.J. et al. (1991) Antisense Res. Dev. 1(3):285-288; Lee, R. et al. (1998) Biochemistry 37(3):900-1010; Pardridge, W.M. et al. (1995) Proc. Natl. Acad. Sci. USA 92(12):5592-5596; and Nielsen, P. E. and Haaima, G. (1997) Chem. Soc. Rev. 96:73-78.) Typically, the binding which results in modulation of expression occurs through hybridization or binding of complementary base pairs. Antisense sequences can also bind to DNA duplexes through specific interactions in the major groove of the double helix.

The polynucleotides of the present invention and fragments thereof can be used as antisense sequences to modify the expression of the polypeptide encoded by mddt. The antisense sequences can be produced ex vivo, such as by using any of the ABI nucleic acid synthesizer series (PE Biosystems) or other automated systems known in the art. Antisense sequences can also be produced biologically, such as by transforming an appropriate host cell with an expression vector containing the sequence of interest. (See, e.g., Agrawal, supra.)

In therapeutic use, any gene delivery system suitable for introduction of the antisense sequences into appropriate target cells can be used. Antisense sequences can be delivered intracellularly in the form of an expression plasmid which, upon transcription, produces a sequence complementary to at least a portion of the cellular sequence encoding the target protein. (See, e.g., Slater, J.E., et al. (1998) J. Allergy Clin. Immunol. 102(3):469-475; and Scanlon, K.J., et al. (1995) 9(13):1288-1296.) Antisense sequences can also be introduced intracellularly through the use of viral vectors, such as retrovirus and adeno-associated virus vectors. (See, e.g., Miller, A.D. (1990) Blood 76:271; Ausubel, F.M. et al. (1995) Current Protocols in Molecular Biology, John Wiley & Sons, New York NY; Uckert, W. and W. Walther (1994) Pharmacol. Ther. 63(3):323-347.) Other gene delivery mechanisms include liposome-derived systems, artificial viral envelopes, and other systems

known in the art. (See, e.g., Rossi, J.J. (1995) Br. Med. Bull. 51(1):217-225; Boado, R.J. et al. (1998) J. Pharm. Sci. 87(11):1308-1315; and Morris, M.C. et al. (1997) Nucleic Acids Res. 25(14):2730-2736.)

Expression

In order to express a biologically active MDDT, the nucleotide sequences encoding MDDT or fragments thereof may be inserted into an appropriate expression vector, i.e., a vector which contains the necessary elements for transcriptional and translational control of the inserted coding sequence in a suitable host. Methods which are well known to those skilled in the art may be used to construct expression vectors containing sequences encoding MDDT and appropriate transcriptional and translational control elements. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination. (See, e.g., Sambrook, supra, Chapters 4, 8, 16, and 17; and Ausubel, supra, Chapters 9, 10, 13, and 16.)

A variety of expression vector/host systems may be utilized to contain and express sequences encoding MDDT. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with viral expression vectors (e.g., baculovirus); plant cell systems transformed with viral expression vectors (e.g., cauliflower mosaic virus, CaMV, or tobacco mosaic virus, TMV) or with bacterial expression vectors (e.g., Ti or pBR322 plasmids); or animal (mammalian) cell systems. (See, e.g., Sambrook, supra; Ausubel, 1995, supra, Van Heeke, G. and S.M. Schuster (1989) J. Biol. Chem. 264:5503-5509; Bitter, G.A. et al. (1987) Methods Enzymol. 153:516-544; Scorer, C.A. et al. (1994) Bio/Technology 12:181-184; Engelhard, E.K. et al. (1994) Proc. Natl. Acad. Sci. USA 91:3224-3227; Sandig, V. et al. (1996) Hum. Gene Ther. 7:1937-1945; Takamatsu, N. (1987) EMBO J. 6:307-311; Coruzzi, G. et al. (1984) EMBO J. 3:1671-1680; Broglie, R. et al. (1984) Science 224:838-843; Winter, J. et al. (1991) Results Probl. Cell Differ. 17:85-105; The McGraw Hill Yearbook of Science and Technology (1992) McGraw Hill, New York NY, pp. 191-196; Logan, J. and T. Shenk (1984) Proc. Natl. Acad. Sci. USA 81:3655-3659; and Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355.) Expression vectors derived from retroviruses, adenoviruses, or herpes or vaccinia viruses, or from various bacterial plasmids, may be used for delivery of nucleotide sequences to the targeted organ, tissue, or cell population. (See, e.g., Di Nicola, M. et al. (1998) Cancer Gen. Ther. 5(6):350-356; Yu, M. et al., (1993) Proc. Natl. Acad. Sci. USA 90(13):6340-6344; Buller, R.M. et al. (1985) Nature 317(6040):813-815; McGregor, D.P. et al. (1994) Mol. Immunol. 31(3):219-226; and Verma, I.M. and N. Somia (1997) Nature 389:239-242.) The invention is not limited by the host cell employed.

For long term production of recombinant proteins in mammalian systems, stable expression of MDDT in cell lines is preferred. For example, sequences encoding MDDT can be transformed into cell lines using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Any number of selection systems may be used to recover transformed cell lines. (See, e.g., Wigler, M. et al. (1977) Cell 11:223-232; Lowy, I. et al. (1980) Cell 22:817-823.; Wigler, M. et al. (1980) Proc. Natl. Acad. Sci. USA 77:3567-3570; Colbere-Garapin, F. et al. (1981) J. Mol. Biol. 150:1-14; Hartman, S.C. and R.C.Mulligan (1988) Proc. Natl. Acad. Sci. USA 85:8047-8051; Rhodes, C.A. (1995) Methods Mol. Biol. 55:121-131.)

Therapeutic Uses of mddt

The mddt of the invention may be used for somatic or germline gene therapy. Gene therapy may be performed to (i) correct a genetic deficiency (e.g., in the cases of severe combined immunodeficiency (SCID)-X1 disease characterized by X-linked inheritance (Cavazzana-Calvo, M. et al. (2000) Science 288:669-672), severe combined immunodeficiency syndrome associated with an inherited adenosine deaminase (ADA) deficiency (Blaese, R.M. et al. (1995) Science 270:475-480; Bordignon, C. et al. (1995) Science 270:470-475), cystic fibrosis (Zabner, J. et al. (1993) Cell 75:207-216; Crystal, R.G. et al. (1995) Hum. Gene Therapy 6:643-666; Crystal, R.G. et al. (1995) Hum. Gene Therapy 6:667-703), thalassemias, familial hypercholesterolemia, and hemophilia resulting from Factor VIII or Factor IX deficiencies (Crystal, R.G. (1995) Science 270:404-410; Verma, I.M. and Somia, N. (1997) Nature 389:239-242)), (ii) express a conditionally lethal gene product (e.g., in the case of cancers which result from unregulated cell proliferation), or (iii) express a protein which affords protection against intracellular parasites (e.g., against human retroviruses, such as human immunodeficiency virus (HIV) (Baltimore, D. (1988) Nature 335:395-396; Poeschla, E. et al. (1996) Proc. Natl. Acad. Sci. USA. 93:11395-11399), hepatitis B or C virus (HBV, HCV); fungal parasites, such as Candida albicans and Paracoccidioides brasiliensis; and protozoan parasites such as Plasmodium falciparum and Trypanosoma cruzi). In the case where a genetic deficiency in mddt expression or regulation causes disease, the expression of mddt from an appropriate population of transduced cells may alleviate the clinical manifestations caused by the genetic deficiency.

In a further embodiment of the invention, diseases or disorders caused by deficiencies in mddt are treated by constructing mammalian expression vectors comprising mddt and introducing these vectors by mechanical means into mddt-deficient cells. Mechanical transfer technologies for use with cells in vivo or ex vitro include (i) direct DNA microinjection into individual cells, (ii) ballistic gold particle delivery, (iii) liposome-mediated transfection, (iv) receptor-mediated gene transfer, and (v) the use of DNA transposons (Morgan, R.A. and Anderson, W.F. (1993) Annu. Rev.

Biochem. 62:191-217; Ivics, Z. (1997) Cell 91:501-510; Boulay, J-L. and Récipon, H. (1998) Curr. Opin. Biotechnol. 9:445-450).

Expression vectors that may be effective for the expression of mddt include, but are not limited to, the PCDNA 3.1, EPITAG, PRCCMV2, PREP, PVAX vectors (Invitrogen, Carlsbad CA), PCMV-SCRIPT, PCMV-TAG, PEGSH/PERV (Stratagene, La Jolla CA), and PTET-OFF, PTET-ON, PTRE2, PTRE2-LUC, PTK-HYG (Clontech, Palo Alto CA). The mddt of the invention may be expressed using (i) a constitutively active promoter, (e.g., from cytomegalovirus (CMV), Rous sarcoma virus (RSV), SV40 virus, thymidine kinase (TK), or β-actin genes), (ii) an inducible promoter (e.g., the tetracycline-regulated promoter (Gossen, M. and Bujard, H. (1992) Proc. Natl. Acad. Sci. U.S.A. 89:5547-5551; Gossen, M. et al., (1995) Science 268:1766-1769; Rossi, F.M.V. and Blau, H.M. (1998) Curr. Opin. Biotechnol. 9:451-456), commercially available in the T-REX plasmid (Invitrogen)); the ecdysone-inducible promoter (available in the plasmids PVGRXR and PIND; Invitrogen); the FK506/rapamycin inducible promoter; or the RU486/mifepristone inducible promoter (Rossi, F.M.V. and Blau, H.M. supra)), or (iii) a tissue-specific promoter or the native promoter of the endogenous gene encoding MDDT from a normal individual.

Commercially available liposome transformation kits (e.g., the PERFECT LIPID TRANSFECTION KIT, available from Invitrogen) allow one with ordinary skill in the art to deliver polynucleotides to target cells in culture and require minimal effort to optimize experimental parameters. In the alternative, transformation is performed using the calcium phosphate method (Graham, F.L. and Eb, A.J. (1973) Virology 52:456-467), or by electroporation (Neumann, E. et al. (1982) EMBO J. 1:841-845). The introduction of DNA to primary cells requires modification of these standardized mammalian transfection protocols.

In another embodiment of the invention, diseases or disorders caused by genetic defects with respect to mddt expression are treated by constructing a retrovirus vector consisting of (i) the mddt of the invention under the control of an independent promoter or the retrovirus long terminal repeat (LTR) promoter, (ii) appropriate RNA packaging signals, and (iii) a Rev-responsive element (RRE) along with additional retrovirus *cis*-acting RNA sequences and coding sequences required for efficient vector propagation. Retrovirus vectors (e.g., PFB and PFBNEO) are commercially available (Stratagene) and are based on published data (Riviere, I. et al. (1995) Proc. Natl. Acad. Sci. U.S.A. 92:6733-6737), incorporated by reference herein. The vector is propagated in an appropriate vector producing cell line (VPCL) that expresses an envelope gene with a tropism for receptors on the target cells or a promiscuous envelope protein such as VSVg (Armentano, D. et al. (1987) J. Virol. 61:1647-1650; Bender, M.A. et al. (1987) J. Virol. 61:1639-1646; Adam, M.A. and Miller, A.D. (1988) J. Virol. 62:3802-3806; Dull, T. et al. (1998) J. Virol. 72:8463-8471; Zufferey, R. et al. (1998) J. Virol. 72:9873-9880). U.S. Patent Number 5,910,434 to Rigg ("Method for obtaining retrovirus packaging

cell lines producing high transducing efficiency retroviral supernatant") discloses a method for obtaining retrovirus packaging cell lines and is hereby incorporated by reference. Propagation of retrovirus vectors, transduction of a population of cells (e.g., $CD4^+$ T-cells), and the return of transduced cells to a patient are procedures well known to persons skilled in the art of gene therapy and have been well documented (Ranga, U. et al. (1997) J. Virol. 71:7020-7029; Bauer, G. et al. (1997) Blood 89:2259-2267; Bonyhadi, M.L. (1997) J. Virol. 71:4707-4716; Ranga, U. et al. (1998) Proc. Natl. Acad. Sci. U.S.A. 95:1201-1206; Su, L. (1997) Blood 89:2283-2290).

In the alternative, an adenovirus-based gene therapy delivery system is used to deliver mddt to cells which have one or more genetic abnormalities with respect to the expression of mddt. The construction and packaging of adenovirus-based vectors are well known to those with ordinary skill in the art. Replication defective adenovirus vectors have proven to be versatile for importing genes encoding immunoregulatory proteins into intact islets in the pancreas (Csete, M.E. et al. (1995) Transplantation 27:263-268). Potentially useful adenoviral vectors are described in U.S. Patent Number 5,707,618 to Armentano ("Adenovirus vectors for gene therapy"), hereby incorporated by reference. For adenoviral vectors, see also Antinozzi, P.A. et al. (1999) Annu. Rev. Nutr. 19:511-544 and Verma, I.M. and Somia, N. (1997) Nature 18:389:239-242, both incorporated by reference herein.

In another alternative, a herpes-based, gene therapy delivery system is used to deliver mddt to target cells which have one or more genetic abnormalities with respect to the expression of mddt. The use of herpes simplex virus (HSV)-based vectors may be especially valuable for introducing mddt to cells of the central nervous system, for which HSV has a tropism. The construction and packaging of herpes-based vectors are well known to those with ordinary skill in the art. A replication-competent herpes simplex virus (HSV) type 1-based vector has been used to deliver a reporter gene to the eyes of primates (Liu, X. et al. (1999) Exp. Eye Res.169:385-395). The construction of a HSV-1 virus vector has also been disclosed in detail in U.S. Patent Number 5,804,413 to DeLuca ("Herpes simplex virus strains for gene transfer"), which is hereby incorporated by reference. U.S. Patent Number 5,804,413 teaches the use of recombinant HSV d92 which consists of a genome containing at least one exogenous gene to be transferred to a cell under the control of the appropriate promoter for purposes including human gene therapy. Also taught by this patent are the construction and use of recombinant HSV strains deleted for ICP4, ICP27 and ICP22. For HSV vectors, see also Goins, W. F. et al. 1999 J. Virol. 73:519-532 and Xu, H. et al., (1994) Dev. Biol. 163:152-161, hereby incorporated by reference. The manipulation of cloned herpesvirus sequences, the generation of recombinant virus following the transfection of multiple plasmids containing different segments of the large herpesvirus genomes, the growth and propagation of herpesvirus, and the infection of cells with herpesvirus are techniques well known to those of ordinary skill in the art.

In another alternative, an alphavirus (positive, single-stranded RNA virus) vector is used to deliver mddt to target cells. The biology of the prototypic alphavirus, Semliki Forest Virus (SFV), has been studied extensively and gene transfer vectors have been based on the SFV genome (Garoff, H. and Li, K-J. (1998) Curr. Opin. Biotech. 9:464-469). During alphavirus RNA replication, a subgenomic RNA is generated that normally encodes the viral capsid proteins. This subgenomic RNA replicates to higher levels than the full-length genomic RNA, resulting in the overproduction of capsid proteins relative to the viral proteins with enzymatic activity (e.g., protease and polymerase). Similarly, inserting mddt into the alphavirus genome in place of the capsid-coding region results in the production of a large number of mddt RNAs and the synthesis of high levels of MDDT in vector transduced cells. While alphavirus infection is typically associated with cell lysis within a few days, the ability to establish a persistent infection in hamster normal kidney cells (BHK-21) with a variant of Sindbis virus (SIN) indicates that the lytic replication of alphaviruses can be altered to suit the needs of the gene therapy application (Dryga, S.A. et al. (1997) Virology 228:74-83). The wide host range of alphaviruses will allow the introduction of MDDT into a variety of cell types. The specific transduction of a subset of cells in a population may require the sorting of cells prior to transduction. The methods of manipulating infectious cDNA clones of alphaviruses, performing alphavirus cDNA and RNA transfections, and performing alphavirus infections, are well known to those with ordinary skill in the art.

Antibodies

Anti-MDDT antibodies may be used to analyze protein expression levels. Such antibodies include, but are not limited to, polyclonal, monoclonal, chimeric, single chain, and Fab fragments. For descriptions of and protocols of antibody technologies, see, e.g., Pound J.D. (1998) Immunochemical Protocols, Humana Press, Totowa, NJ.

The amino acid sequence encoded by the mddt of the Sequence Listing may be analyzed by appropriate software (e.g., LASERGENE NAVIGATOR software, DNASTAR) to determine regions of high immunogenicity. The optimal sequences for immunization are selected from the C-terminus, the N-terminus, and those intervening, hydrophilic regions of the polypeptide which are likely to be exposed to the external environment when the polypeptide is in its natural conformation. Analysis used to select appropriate epitopes is also described by Ausubel (1997, supra, Chapter 11.7). Peptides used for antibody induction do not need to have biological activity; however, they must be antigenic. Peptides used to induce specific antibodies may have an amino acid sequence consisting of at five amino acids, preferably at least 10 amino acids, and most preferably 15 amino acids. A peptide which mimics an antigenic fragment of the natural polypeptide may be fused with another protein such as keyhole limpet cyanin (KLH; Sigma, St. Louis MO) for antibody production. A peptide

encompassing an antigenic region may be expressed from an mddt, synthesized as described above, or purified from human cells.

Procedures well known in the art may be used for the production of antibodies. Various hosts including mice, goats, and rabbits, may be immunized by injection with a peptide. Depending on the host species, various adjuvants may be used to increase immunological response.

In one procedure, peptides about 15 residues in length may be synthesized using an ABI 431A peptide synthesizer (PE Biosystems) using fmoc-chemistry and coupled to KLH (Sigma) by reaction with M-maleimidobenzoyl-N-hydroxysuccinimide ester (Ausubel, 1995, supra). Rabbits are immunized with the peptide-KLH complex in complete Freund's adjuvant. The resulting antisera are tested for antipeptide activity by binding the peptide to plastic, blocking with 1% bovine serum albumin (BSA), reacting with rabbit antisera, washing, and reacting with radioiodinated goat anti-rabbit IgG. Antisera with antipeptide activity are tested for anti-MDDT activity using protocols well known in the art, including ELISA, radioimmunoassay (RIA), and immunoblotting.

In another procedure, isolated and purified peptide may be used to immunize mice (about 100 µg of peptide) or rabbits (about 1 mg of peptide). Subsequently, the peptide is radioiodinated and used to screen the immunized animals' B-lymphocytes for production of antipeptide antibodies. Positive cells are then used to produce hybridomas using standard techniques. About 20 mg of peptide is sufficient for labeling and screening several thousand clones. Hybridomas of interest are detected by screening with radioiodinated peptide to identify those fusions producing peptide-specific monoclonal antibody. In a typical protocol, wells of a multi-well plate (FAST, Becton-Dickinson, Palo Alto, CA) are coated with affinity-purified, specific rabbit-anti-mouse (or suitable anti-species IgG) antibodies at 10 mg/ml. The coated wells are blocked with 1% BSA and washed and exposed to supernatants from hybridomas. After incubation, the wells are exposed to radiolabeled peptide at 1 mg/ml.

Clones producing antibodies bind a quantity of labeled peptide that is detectable above background. Such clones are expanded and subjected to 2 cycles of cloning. Cloned hybridomas are injected into pristane-treated mice to produce ascites, and monoclonal antibody is purified from the ascitic fluid by affinity chromatography on protein A (Amersham Pharmacia Biotech). Several procedures for the production of monoclonal antibodies, including in vitro production, are described in Pound (supra). Monoclonal antibodies with antipeptide activity are tested for anti-MDDT activity using protocols well known in the art, including ELISA, RIA, and immunoblotting.

Antibody fragments containing specific binding sites for an epitope may also be generated. For example, such fragments include, but are not limited to, the F(ab')2 fragments produced by pepsin digestion of the antibody molecule, and the Fab fragments generated by reducing the disulfide bridges of the F(ab')2 fragments. Alternatively, construction of Fab expression libraries in filamentous

bacteriophage allows rapid and easy identification of monoclonal fragments with desired specificity (Pound, supra, Chaps. 45-47). Antibodies generated against polypeptide encoded by mddt can be used to purify and characterize full-length MDDT protein and its activity, binding partners, etc.

Assays Using Antibodies

Anti-MDDT antibodies may be used in assays to quantify the amount of MDDT found in a particular human cell. Such assays include methods utilizing the antibody and a label to detect expression level under normal or disease conditions. The peptides and antibodies of the invention may be used with or without modification or labeled by joining them, either covalently or noncovalently, with a reporter molecule.

Protocols for detecting and measuring protein expression using either polyclonal or monoclonal antibodies are well known in the art. Examples include ELISA, RIA, and fluorescent activated cell sorting (FACS). Such immunoassays typically involve the formation of complexes between the MDDT and its specific antibody and the measurement of such complexes. These and other assays are described in Pound (supra).

Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. The following preferred specific embodiments are, therefore, to be construed as merely illustrative, and not limitative of the remainder of the disclosure in any way whatsoever.

The disclosures of all patents, applications, and publications mentioned above and below, in particular U.S. Provisional Application No. 60/137,412, filed June 3, 1999, U.S. Provisional Application No. 60/147,542, filed August 5, 1999, U.S. Provisional Application No. 60/147,501, filed August 5, 1999, U.S. Provisional Application No. 60/147,500, filed August 5, 1999 are hereby expressly incorporated by reference.

EXAMPLES

I. Construction of cDNA Libraries

RNA was purchased from CLONTECH Laboratories, Inc. (Palo Alto CA) or isolated from various tissues. Some tissues were homogenized and lysed in guanidinium isothiocyanate, while others were homogenized and lysed in phenol or in a suitable mixture of denaturants, such as TRIZOL (Life Technologies), a monophasic solution of phenol and guanidine isothiocyanate. The resulting lysates were centrifuged over CsCl cushions or extracted with chloroform. RNA was precipitated with either isopropanol or sodium acetate and ethanol, or by other routine methods.

Phenol extraction and precipitation of RNA were repeated as necessary to increase RNA purity. In most cases, RNA was treated with DNase. For most libraries, poly(A+) RNA was isolated

using oligo d(T)-coupled paramagnetic particles (Promega Corporation (Promega), Madison WI), OLIGOTEX latex particles (QIAGEN, Inc. (QIAGEN), Valencia CA), or an OLIGOTEX mRNA purification kit (QIAGEN). Alternatively, RNA was isolated directly from tissue lysates using other RNA isolation kits, e.g., the POLY(A)PURE mRNA purification kit (Ambion, Inc., Austin TX).

In some cases, Stratagene was provided with RNA and constructed the corresponding cDNA libraries. Otherwise, cDNA was synthesized and cDNA libraries were constructed with the UNIZAP vector system (Stratagene Cloning Systems, Inc. (Stratagene), La Jolla CA) or SUPERSCRIPT plasmid system (Life Technologies), using the recommended procedures or similar methods known in the art. (See, e.g., Ausubel, 1997, supra, Chapters 5.1 through 6.6.) Reverse transcription was initiated using oligo d(T) or random primers. Synthetic oligonucleotide adapters were ligated to double stranded cDNA, and the cDNA was digested with the appropriate restriction enzyme or enzymes. For most libraries, the cDNA was size-selected (300-1000 bp) using SEPHACRYL S1000, SEPHAROSE CL2B, or SEPHAROSE CL4B column chromatography (Amersham Pharmacia Biotech) or preparative agarose gel electrophoresis. cDNAs were ligated into compatible restriction enzyme sites of the polylinker of a suitable plasmid, e.g., PBLUESCRIPT plasmid (Stratagene), pSPORT1 plasmid (Life Technologies), or pINCY (Incyte). Recombinant plasmids were transformed into competent E. coli cells including XL1-Blue, XL1-BlueMRF, or SOLR from Stratagene or DH5α, DH10B, or ElectroMAX DH10B from Life Technologies.

II. Isolation of cDNA Clones

Plasmids were recovered from host cells by in vivo excision using the UNIZAP vector system (Stratagene) or by cell lysis. Plasmids were purified using at least one of the following: the Magic or WIZARD Minipreps DNA purification system (Promega); the AGTC Miniprep purification kit (Edge BioSystems, Gaithersburg MD); and the QIAWELL 8, QIAWELL 8 Plus, and QIAWELL 8 Ultra plasmid purification systems or the R.E.A.L. PREP 96 plasmid purification kit (QIAGEN). Following precipitation, plasmids were resuspended in 0.1 ml of distilled water and stored, with or without lyophilization, at 4°C.

Alternatively, plasmid DNA was amplified from host cell lysates using direct link PCR in a high-throughput format. (Rao, V.B. (1994) Anal. Biochem. 216:1-14.) Host cell lysis and thermal cycling steps were carried out in a single reaction mixture. Samples were processed and stored in 384-well plates, and the concentration of amplified plasmid DNA was quantified fluorometrically using PICOGREEN dye (Molecular Probes, Inc. (Molecular Probes), Eugene OR) and a FLUOROSKAN II fluorescence scanner (Labsystems Oy, Helsinki, Finland).

III. Sequencing and Analysis

cDNA sequencing reactions were processed using standard methods or high-throughput instrumentation such as the ABI CATALYST 800 thermal cycler (PE Biosystems) or the PTC-200 thermal cycler (MJ Research) in conjunction with the HYDRA microdispenser (Robbins Scientific Corp., Sunnyvale CA) or the MICROLAB 2200 liquid transfer system (Hamilton). cDNA sequencing reactions were prepared using reagents provided by Amersham Pharmacia Biotech or supplied in ABI sequencing kits such as the ABI PRISM BIGDYE Terminator cycle sequencing ready reaction kit (PE Biosystems). Electrophoretic separation of cDNA sequencing reactions and detection of labeled polynucleotides were carried out using the MEGABACE 1000 DNA sequencing system (Molecular Dynamics); the ABI PRISM 373 or 377 sequencing system (PE Biosystems) in conjunction with standard ABI protocols and base calling software; or other sequence analysis systems known in the art. Reading frames within the cDNA sequences were identified using standard methods (reviewed in Ausubel, 1997, supra, Chapter 7.7). Some of the cDNA sequences were selected for extension using the techniques disclosed in Example VIII.

IV. Assembly and Analysis of Sequences

Component sequences from chromatograms were subject to PHRED analysis and assigned a quality score. The sequences having at least a required quality score were subject to various pre-processing editing pathways to eliminate, e.g., low quality 3' ends, vector and linker sequences, polyA tails, Alu repeats, mitochondrial and ribosomal sequences, bacterial contamination sequences, and sequences smaller than 50 base pairs. In particular, low-information sequences and repetitive elements (e.g., dinucleotide repeats, Alu repeats, etc.) were replaced by "n's", or masked, to prevent spurious matches.

Processed sequences were then subject to assembly procedures in which the sequences were assigned to gene bins (bins). Each sequence could only belong to one bin. Sequences in each gene bin were assembled to produce consensus sequences (templates). Subsequent new sequences were added to existing bins using BLASTn (v.1.4 WashU) and CROSSMATCH. Candidate pairs were identified as all BLAST hits having a quality score greater than or equal to 150. Alignments of at least 82% local identity were accepted into the bin. The component sequences from each bin were assembled using a version of PHRAP. Bins with several overlapping component sequences were assembled using DEEP PHRAP. The orientation (sense or antisense) of each assembled template was determined based on the number and orientation of its component sequences. Template sequences as disclosed in the sequence listing correspond to sense strand sequences (the "forward" reading frames), to the best determination. The complementary (antisense) strands are inherently disclosed

herein. The component sequences which were used to assemble each template consensus sequence are listed in Table 4, along with their positions along the template nucleotide sequences.

Bins were compared against each other and those having local similarity of at least 82% were combined and reassembled. Reassembled bins having templates of insufficient overlap (less than 95% local identity) were re-split. Assembled templates were also subject to analysis by STITCHER/EXON MAPPER algorithms which analyze the probabilities of the presence of splice variants, alternatively spliced exons, splice junctions, differential expression of alternative spliced genes across tissue types or disease states, etc. These resulting bins were subject to several rounds of the above assembly procedures.

Once gene bins were generated based upon sequence alignments, bins were clone joined based upon clone information. If the 5' sequence of one clone was present in one bin and the 3' sequence from the same clone was present in a different bin, it was likely that the two bins actually belonged together in a single bin. The resulting combined bins underwent assembly procedures to regenerate the consensus sequences.

The final assembled templates were subsequently annotated using the following procedure. Template sequences were analyzed using BLASTn (v2.0, NCBI) versus gbpri (GenBank version 116). "Hits" were defined as an exact match having from 95% local identity over 200 base pairs through 100% local identity over 100 base pairs, or a homolog match having an E-value, i.e. a probability score, of $\leq 1 \times 10^{-8}$. The hits were subject to frameshift FASTx versus GENPEPT (GenBank version 116). (See Table 5). In this analysis, a homolog match was defined as having an E-value of $\leq 1 \times 10^{-8}$. The assembly method used above was described in "System and Methods for Analyzing Biomolecular Sequences," U.S.S.N. 09/276,534, filed March 25, 1999, and the LIFESEQ Gold user manual (Incyte) both incorporated by reference herein.

Following assembly, template sequences were subjected to motif, BLAST, and functional analyses, and categorized in protein hierarchies using methods described in, e.g., "Database System Employing Protein Function Hierarchies for Viewing Biomolecular Sequence Data," U.S.S.N. 08/812,290, filed March 6, 1997; "Relational Database for Storing Biomolecule Information," U.S.S.N. 08/947,845, filed October 9, 1997; "Project-Based Full-Length Biomolecular Sequence Database," U.S.S.N. 08/811,758, filed March 6, 1997; and "Relational Database and System for Storing Information Relating to Biomolecular Sequences," U.S.S.N. 09/034,807, filed March 4, 1998, all of which are incorporated by reference herein.

The template sequences were further analyzed by translating each template in all three forward reading frames and searching each translation against the Pfam database of hidden Markov model-based protein families and domains using the HMMER software package (available to the public from Washington University School of Medicine, St. Louis MO). Regions of templates which,

when translated, contain similarity to Pfam consensus sequences are reported in Table 2, along with descriptions of Pfam protein domains and families. Only those Pfam hits with an E-value of $\leq 1 \times 10^{-3}$ are reported. (See also World Wide Web site http://pfam.wustl.edu/ for detailed descriptions of Pfam protein domains and families.)

Additionally, the template sequences were translated in all three forward reading frames, and each translation was searched against hidden Markov models for signal peptide and transmembrane domains using the HMMER software package. Construction of hidden Markov models and their usage in sequence analysis has been described. (See, for example, Eddy, S.R. (1996) Curr. Opin. Str. Biol. 6:361-365.) Regions of templates which, when translated, contain similarity to signal peptide or transmembrane domain consensus sequences are reported in Table 3. Only those signal peptide or transmembrane hits with a cutoff score of 11 bits or greater are reported. A cutoff score of 11 bits or greater corresponds to at least about 91-94% true-positives in signal peptide prediction, and at least about 75% true-positives in transmembrane domain prediction.

The results of HMMER analysis as reported in Tables 2 and 3 may support the results of BLAST analysis as reported in Table 1 or may suggest alternative or additional properties of template-encoded polypeptides not previously uncovered by BLAST or other analyses.

Template sequences are further analyzed using the bioinformatics tools listed in Table 5, or using sequence analysis software known in the art such as MACDNASIS PRO software (Hitachi Software Engineering, South San Francisco CA) and LASERGENE software (DNASTAR). Template sequences may be further queried against public databases such as the GenBank rodent, mammalian, vertebrate, prokaryote, and eukaryote databases.

V. Analysis of Polynucleotide Expression

Northern analysis is a laboratory technique used to detect the presence of a transcript of a gene and involves the hybridization of a labeled nucleotide sequence to a membrane on which RNAs from a particular cell type or tissue have been bound. (See, e.g., Sambrook, supra, ch. 7; Ausubel, 1995, supra, ch. 4 and 16.)

Analogous computer techniques applying BLAST were used to search for identical or related molecules in cDNA databases such as GenBank or LIFESEQ (Incyte Pharmaceuticals). This analysis is much faster than multiple membrane-based hybridizations. In addition, the sensitivity of the computer search can be modified to determine whether any particular match is categorized as exact or similar. The basis of the search is the product score, which is defined as:

$$\frac{\text{BLAST Score} \times \text{Percent Identity}}{5 \times \text{minimum}\{\text{length(Seq. 1)}, \text{length(Seq. 2)}\}}$$

The product score takes into account both the degree of similarity between two sequences and the length of the sequence match. The product score is a normalized value between 0 and 100, and is calculated as follows: the BLAST score is multiplied by the percent nucleotide identity and the product is divided by (5 times the length of the shorter of the two sequences). The BLAST score is calculated by assigning a score of +5 for every base that matches in a high-scoring segment pair (HSP), and -4 for every mismatch. Two sequences may share more than one HSP (separated by gaps). If there is more than one HSP, then the pair with the highest BLAST score is used to calculate the product score. The product score represents a balance between fractional overlap and quality in a BLAST alignment. For example, a product score of 100 is produced only for 100% identity over the entire length of the shorter of the two sequences being compared. A product score of 70 is produced either by 100% identity and 70% overlap at one end, or by 88% identity and 100% overlap at the other. A product score of 50 is produced either by 100% identity and 50% overlap at one end, or 79% identity and 100% overlap.

VI. Tissue Distribution Profiling

A tissue distribution profile is determined for each template by compiling the cDNA library tissue classifications of its component cDNA sequences. Each component sequence, is derived from a cDNA library constructed from a human tissue. Each human tissue is classified into one of the following categories: cardiovascular system; connective tissue; digestive system; embryonic structures; endocrine system; exocrine glands; genitalia, female; genitalia, male; germ cells; hemic and immune system; liver; musculoskeletal system; nervous system; pancreas; respiratory system; sense organs; skin; stomatognathic system; unclassified/mixed; or urinary tract. Template sequences, component sequences, and cDNA library/tissue information are found in the LIFESEQ GOLD database (Incyte Genomics, Palo Alto CA).

VII. Transcript Image Analysis

Transcript images are generated as described in Seilhamer et al., "Comparative Gene Transcript Analysis," U.S. Patent Number 5,840,484, incorporated herein by reference.

VIII. Extension of Polynucleotide Sequences and Isolation of a Full-length cDNA

Oligonucleotide primers designed using an mddt of the Sequence Listing are used to extend the nucleic acid sequence. One primer is synthesized to initiate 5' extension of the template, and the other primer, to initiate 3' extension of the template. The initial primers may be designed using OLIGO 4.06 software (National Biosciences, Inc. (National Biosciences), Plymouth MN), or another appropriate program, to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or

r re, and to anneal to the target sequence at temperatures of about 68 °C to about 72 °C. Any stretch c. nucleotides which would result in hairpin structures and primer-primer dimerizations are avoided. Selected human cDNA libraries are used to extend the sequence. If more than one extension is necessary or desired, additional or nested sets of primers are designed.

High fidelity amplification is obtained by PCR using methods well known in the art. PCR is performed in 96-well plates using the PTC-200 thermal cycler (MJ Research). The reaction mix contains DNA template, 200 nmol of each primer, reaction buffer containing Mg^{2+}, $(NH_4)_2SO_4$, and ß-mercaptoethanol, Taq DNA polymerase (Amersham Pharmacia Biotech), ELONGASE enzyme (Life Technologies), and Pfu DNA polymerase (Stratagene), with the following parameters for primer pair PCI A and PCI B: Step 1: 94 °C, 3 min; Step 2: 94 °C, 15 sec; Step 3: 60 °C, 1 min; Step 4: 68 °C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68 °C, 5 min; Step 7: storage at 4 °C. In the alternative, the parameters for primer pair T7 and SK+ are as follows: Step 1: 94 °C, 3 min; Step 2: 94 °C, 15 sec; Step 3: 57 °C, 1 min; Step 4: 68 °C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68 °C, 5 min; Step 7: storage at 4 °C.

The concentration of DNA in each well is determined by dispensing 100 µl PICOGREEN quantitation reagent (0.25% (v/v); Molecular Probes) dissolved in 1X Tris-EDTA (TE) and 0.5 µl of undiluted PCR product into each well of an opaque fluorimeter plate (Corning Incorporated (Corning), Corning NY), allowing the DNA to bind to the reagent. The plate is scanned in a FLUOROSKAN II (Labsystems Oy) to measure the fluorescence of the sample and to quantify the concentration of DNA. A 5 µl to 10 µl aliquot of the reaction mixture is analyzed by electrophoresis on a 1 % agarose mini-gel to determine which reactions are successful in extending the sequence.

The extended nucleotides are desalted and concentrated, transferred to 384-well plates, digested with CviJI cholera virus endonuclease (Molecular Biology Research, Madison WI), and sonicated or sheared prior to religation into pUC 18 vector (Amersham Pharmacia Biotech). For shotgun sequencing, the digested nucleotides are separated on low concentration (0.6 to 0.8%) agarose gels, fragments are excised, and agar digested with AGAR ACE (Promega). Extended clones are religated using T4 ligase (New England Biolabs, Inc., Beverly MA) into pUC 18 vector (Amersham Pharmacia Biotech), treated with Pfu DNA polymerase (Stratagene) to fill-in restriction site overhangs, and transfected into competent E. coli cells. Transformed cells are selected on antibiotic-containing media, individual colonies are picked and cultured overnight at 37 °C in 384-well plates in LB/2x carbenicillin liquid media.

The cells are lysed, and DNA is amplified by PCR using Taq DNA polymerase (Amersham Pharmacia Biotech) and Pfu DNA polymerase (Stratagene) with the following parameters: Step 1: 94 °C, 3 min; Step 2: 94 °C, 15 sec; Step 3: 60 °C, 1 min; Step 4: 72 °C, 2 min; Step 5: steps 2, 3, and 4 repeated 29 times; Step 6: 72 °C, 5 min; Step 7: storage at 4 °C. DNA is quantified by PICOGREEN

reagent (Molecular Probes) as described above. Samples with low DNA recoveries are reamplified using the same conditions as described above. Samples are diluted with 20% dimethysulfoxide (1:2, v/v), and sequenced using DYENAMIC energy transfer sequencing primers and the DYENAMIC DIRECT kit (Amersham Pharmacia Biotech) or the ABI PRISM BIGDYE Terminator cycle sequencing ready reaction kit (PE Biosystems).

In like manner, the mddt is used to obtain regulatory sequences (promoters, introns, and enhancers) using the procedure above, oligonucleotides designed for such extension, and an appropriate genomic library.

IX. Labeling of Probes and Southern Hybridization Analyses

Hybridization probes derived from the mddt of the Sequence Listing are employed for screening cDNAs, mRNAs, or genomic DNA. The labeling of probe nucleotides between 100 and 1000 nucleotides in length is specifically described, but essentially the same procedure may be used with larger cDNA fragments. Probe sequences are labeled at room temperature for 30 minutes using a T4 polynucleotide kinase, γ^{32}P-ATP, and 0.5X One-Phor-All Plus (Amersham Pharmacia Biotech) buffer and purified using a ProbeQuant G-50 Microcolumn (Amersham Pharmacia Biotech). The probe mixture is diluted to 10^7 dpm/µg/ml hybridization buffer and used in a typical membrane-based hybridization analysis.

The DNA is digested with a restriction endonuclease such as Eco RV and is electrophoresed through a 0.7% agarose gel. The DNA fragments are transferred from the agarose to nylon membrane (NYTRAN Plus, Schleicher & Schuell, Inc., Keene NH) using procedures specified by the manufacturer of the membrane. Prehybridization is carried out for three or more hours at 68°C, and hybridization is carried out overnight at 68°C. To remove non-specific signals, blots are sequentially washed at room temperature under increasingly stringent conditions, up to 0.1x saline sodium citrate (SSC) and 0.5% sodium dodecyl sulfate. After the blots are placed in a PHOSPHORIMAGER cassette (Molecular Dynamics) or are exposed to autoradiography film, hybridization patterns of standard and experimental lanes are compared. Essentially the same procedure is employed when screening RNA.

X. Chromosome Mapping of mddt

The cDNA sequences which were used to assemble SEQ ID NO:1-14 are compared with sequences from the Incyte LIFESEQ database and public domain databases using BLAST and other implementations of the Smith-Waterman algorithm. Sequences from these databases that match SEQ ID NO:1-14 are assembled into clusters of contiguous and overlapping sequences using assembly algorithms such as PHRAP (Table 5). Radiation hybrid and genetic mapping data available from

public resources such as the Stanford Human Genome Center (SHGC), Whitehead Institute for Genome Research (WIGR), and Généthon are used to determine if any of the clustered sequences have been previously mapped. Inclusion of a mapped sequence in a cluster will result in the assignment of all sequences of that cluster, including its particular SEQ ID NO:, to that map location. The genetic map locations of SEQ ID NO:1-14 are described as ranges, or intervals, of human chromosomes. The map position of an interval, in centiMorgans, is measured relative to the terminus of the chromosome's p-arm. (The centiMorgan (cM) is a unit of measurement based on recombination frequencies between chromosomal markers. On average, 1 cM is roughly equivalent to 1 megabase (Mb) of DNA in humans, although this can vary widely due to hot and cold spots of recombination.) The cM distances are based on genetic markers mapped by Généthon which provide boundaries for radiation hybrid markers whose sequences were included in each of the clusters.

XI. Microarray Analysis

Probe Preparation from Tissue or Cell Samples

Total RNA is isolated from tissue samples using the guanidinium thiocyanate method and polyA$^+$ RNA is purified using the oligo (dT) cellulose method. Each polyA$^+$ RNA sample is reverse transcribed using MMLV reverse-transcriptase, 0.05 pg/µl oligo-dT primer (21mer), 1X first strand buffer, 0.03 units/µl RNase inhibitor, 500 µM dATP, 500 µM dGTP, 500 µM dTTP, 40 µM dCTP, 40 µM dCTP-Cy3 (BDS) or dCTP-Cy5 (Amersham Pharmacia Biotech). The reverse transcription reaction is performed in a 25 ml volume containing 200 ng polyA$^+$ RNA with GEMBRIGHT kits (Incyte). Specific control polyA$^+$ RNAs are synthesized by in vitro transcription from non-coding yeast genomic DNA (W. Lei, unpublished). As quantitative controls, the control mRNAs at 0.002 ng, 0.02 ng, 0.2 ng, and 2 ng are diluted into reverse transcription reaction at ratios of 1:100,000, 1:10,000, 1:1000, 1:100 (w/w) to sample mRNA respectively. The control mRNAs are diluted into reverse transcription reaction at ratios of 1:3, 3:1, 1:10, 10:1, 1:25, 25:1 (w/w) to sample mRNA differential expression patterns. After incubation at 37° C for 2 hr, each reaction sample (one with Cy3 and another with Cy5 labeling) is treated with 2.5 ml of 0.5M sodium hydroxide and incubated for 20 minutes at 85° C to the stop the reaction and degrade the RNA. Probes are purified using two successive CHROMA SPIN 30 gel filtration spin columns (CLONTECH Laboratories, Inc. (CLONTECH), Palo Alto CA) and after combining, both reaction samples are ethanol precipitated using 1 ml of glycogen (1 mg/ml), 60 ml sodium acetate, and 300 ml of 100% ethanol. The probe is then dried to completion using a SpeedVAC (Savant Instruments Inc., Holbrook NY) and resuspended in 14 µl 5X SSC/0.2% SDS.

Microarray Preparation

Sequences of the present invention are used to generate array elements. Each array element is amplified from bacterial cells containing vectors with cloned cDNA inserts. PCR amplification uses primers complementary to the vector sequences flanking the cDNA insert. Array elements are amplified in thirty cycles of PCR from an initial quantity of 1-2 ng to a final quantity greater than 5 µg. Amplified array elements are then purified using SEPHACRYL-400 (Amersham Pharmacia Biotech).

Purified array elements are immobilized on polymer-coated glass slides. Glass microscope slides (Corning) are cleaned by ultrasound in 0.1% SDS and acetone, with extensive distilled water washes between and after treatments. Glass slides are etched in 4% hydrofluoric acid (VWR Scientific Products Corporation (VWR), West Chester, PA), washed extensively in distilled water, and coated with 0.05% aminopropyl silane (Sigma) in 95% ethanol. Coated slides are cured in a 110°C oven.

Array elements are applied to the coated glass substrate using a procedure described in US Patent No. 5,807,522, incorporated herein by reference. 1 µl of the array element DNA, at an average concentration of 100 ng/µl, is loaded into the open capillary printing element by a high-speed robotic apparatus. The apparatus then deposits about 5 nl of array element sample per slide.

Microarrays are UV-crosslinked using a STRATALINKER UV-crosslinker (Stratagene). Microarrays are washed at room temperature once in 0.2% SDS and three times in distilled water. Non-specific binding sites are blocked by incubation of microarrays in 0.2% casein in phosphate buffered saline (PBS) (Tropix, Inc., Bedford, MA) for 30 minutes at 60° C followed by washes in 0.2% SDS and distilled water as before.

Hybridization

Hybridization reactions contain 9 µl of probe mixture consisting of 0.2 µg each of Cy3 and Cy5 labeled cDNA synthesis products in 5X SSC, 0.2% SDS hybridization buffer. The probe mixture is heated to 65° C for 5 minutes and is aliquoted onto the microarray surface and covered with an 1.8 cm^2 coverslip. The arrays are transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber is kept at 100% humidity internally by the addition of 140 µl of 5x SSC in a corner of the chamber. The chamber containing the arrays is incubated for about 6.5 hours at 60° C. The arrays are washed for 10 min at 45° C in a first wash buffer (1X SSC, 0.1% SDS), three times for 10 minutes each at 45° C in a second wash buffer (0.1X SSC), and dried.

Detection

Reporter-labeled hybridization complexes are detected with a microscope equipped with an

Innova 70 mixed gas 10 W laser (Coherent, Inc., Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light is focused on the array using a 20X microscope objective (Nikon, Inc., Melville NY). The slide containing the array is placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective. The 1.8 cm x 1.8 cm array used in the present example is scanned with a resolution of 20 micrometers.

In two separate scans, a mixed gas multiline laser excites the two fluorophores sequentially. Emitted light is split, based on wavelength, into two photomultiplier tube detectors (PMT R1477, Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Appropriate filters positioned between the array and the photomultiplier tubes are used to filter the signals. The emission maxima of the fluorophores used are 565 nm for Cy3 and 650 nm for Cy5. Each array is typically scanned twice, one scan per fluorophore using the appropriate filters at the laser source, although the apparatus is capable of recording the spectra from both fluorophores simultaneously.

The sensitivity of the scans is typically calibrated using the signal intensity generated by a cDNA control species added to the probe mix at a known concentration. A specific location on the array contains a complementary DNA sequence, allowing the intensity of the signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000. When two probes from different sources (e.g., representing test and control cells), each labeled with a different fluorophore, are hybridized to a single array for the purpose of identifying genes that are differentially expressed, the calibration is done by labeling samples of the calibrating cDNA with the two fluorophores and adding identical amounts of each to the hybridization mixture.

The output of the photomultiplier tube is digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Inc., Norwood, MA) installed in an IBM-compatible PC computer. The digitized data are displayed as an image where the signal intensity is mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data is also analyzed quantitatively. Where two different fluorophores are excited and measured simultaneously, the data are first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using each fluorophore's emission spectrum.

A grid is superimposed over the fluorescence signal image such that the signal from each spot is centered in each element of the grid. The fluorescence signal within each element is then integrated to obtain a numerical value corresponding to the average intensity of the signal. The software used for signal analysis is the GEMTOOLS gene expression analysis program (Incyte).

XII. Complementary Nucleic Acids

Sequences complementary to the mddt are used to detect, decrease, or inhibit expression of

the naturally occurring nucleotide. The use of oligonucleotides comprising from about 15 to 30 base pairs is typical in the art. However, smaller or larger sequence fragments can also be used. Appropriate oligonucleotides are designed from the mddt using OLIGO 4.06 software (National Biosciences) or other appropriate programs and are synthesized using methods standard in the art or ordered from a commercial supplier. To inhibit transcription, a complementary oligonucleotide is designed from the most unique 5' sequence and used to prevent transcription factor binding to the promoter sequence. To inhibit translation, a complementary oligonucleotide is designed to prevent ribosomal binding and processing of the transcript.

XIII. Expression of MDDT

Expression and purification of MDDT is accomplished using bacterial or virus-based expression systems. For expression of MDDT in bacteria, cDNA is subcloned into an appropriate vector containing an antibiotic resistance gene and an inducible promoter that directs high levels of cDNA transcription. Examples of such promoters include, but are not limited to, the *trp-lac* (*tac*) hybrid promoter and the T5 or T7 bacteriophage promoter in conjunction with the *lac* operator regulatory element. Recombinant vectors are transformed into suitable bacterial hosts, e.g., BL21(DE3). Antibiotic resistant bacteria express MDDT upon induction with isopropyl beta-D-thiogalactopyranoside (IPTG). Expression of MDDT in eukaryotic cells is achieved by infecting insect or mammalian cell lines with recombinant Autographica californica nuclear polyhedrosis virus (AcMNPV), commonly known as baculovirus. The nonessential polyhedrin gene of baculovirus is replaced with cDNA encoding MDDT by either homologous recombination or bacterial-mediated transposition involving transfer plasmid intermediates. Viral infectivity is maintained and the strong polyhedrin promoter drives high levels of cDNA transcription. Recombinant baculovirus is used to infect Spodoptera frugiperda (Sf9) insect cells in most cases, or human hepatocytes, in some cases. Infection of the latter requires additional genetic modifications to baculovirus. (See e.g., Engelhard, supra; and Sandig, supra.)

In most expression systems, MDDT is synthesized as a fusion protein with, e.g., glutathione S-transferase (GST) or a peptide epitope tag, such as FLAG or 6-His, permitting rapid, single-step, affinity-based purification of recombinant fusion protein from crude cell lysates. GST, a 26-kilodalton enzyme from Schistosoma japonicum, enables the purification of fusion proteins on immobilized glutathione under conditions that maintain protein activity and antigenicity (Amersham Pharmacia Biotech). Following purification, the GST moiety can be proteolytically cleaved from MDDT at specifically engineered sites. FLAG, an 8-amino acid peptide, enables immunoaffinity purification using commercially available monoclonal and polyclonal anti-FLAG antibodies (Eastman Kodak Company, Rochester NY). 6-His, a stretch of six consecutive histidine residues, enables

purification on metal-chelate resins (QIAGEN). Methods for protein expression and purification are discussed in Ausubel (1995, supra, Chapters 10 and 16). Purified MDDT obtained by these methods can be used directly in the following activity assay.

XIV. Demonstration of MDDT Activity

MDDT, or biologically active fragments thereof, are labeled with ^{125}I Bolton-Hunter reagent. (See, e.g., Bolton, A.E. and W.M. Hunter (1973) Biochem. J. 133:529-539.) Candidate molecules previously arrayed in the wells of a multi-well plate are incubated with the labeled MDDT, washed, and any wells with labeled MDDT complex are assayed. Data obtained using different concentrations of MDDT are used to calculate values for the number, affinity, and association of MDDT with the candidate molecules.

Alternatively, molecules interacting with MDDT are analyzed using the yeast two-hybrid system as described in Fields, S. and O. Song (1989) Nature 340:245-246, or using commercially available kits based on the two-hybrid system, such as the MATCHMAKER system (CLONTECH).

MDDT may also be used in the PATHCALLING process (CuraGen Corp., New Haven CT) which employs the yeast two-hybrid system in a high-throughput manner to determine all interactions between the proteins encoded by two large libraries of genes (Nandabalan, K. et al. (2000) U.S. Patent No. 6,057,101).

XV. Functional Assays

MDDT function is assessed by expressing mddt at physiologically elevated levels in mammalian cell culture systems. cDNA is subcloned into a mammalian expression vector containing a strong promoter that drives high levels of cDNA expression. Vectors of choice include pCMV SPORT (Life Technologies) and pCR3.1 (Invitrogen Corporation, Carlsbad CA), both of which contain the cytomegalovirus promoter. 5-10 µg of recombinant vector are transiently transfected into a human cell line, preferably of endothelial or hematopoietic origin, using either liposome formulations or electroporation. 1-2 µg of an additional plasmid containing sequences encoding a marker protein are co-transfected.

Expression of a marker protein provides a means to distinguish transfected cells from nontransfected cells and is a reliable predictor of cDNA expression from the recombinant vector. Marker proteins of choice include, e.g., Green Fluorescent Protein (GFP; CLONTECH), CD64, or a CD64-GFP fusion protein. Flow cytometry (FCM), an automated laser optics-based technique, is used to identify transfected cells expressing GFP or CD64-GFP and to evaluate the apoptotic state of the cells and other cellular properties.

FCM detects and quantifies the uptake of fluorescent molecules that diagnose events preceding or coincident with cell death. These events include changes in nuclear DNA content as measured by staining of DNA with propidium iodide; changes in cell size and granularity as measured by forward light scatter and 90 degree side light scatter; down-regulation of DNA synthesis as measured by decrease in bromodeoxyuridine uptake; alterations in expression of cell surface and intracellular proteins as measured by reactivity with specific antibodies; and alterations in plasma membrane composition as measured by the binding of fluorescein-conjugated Annexin V protein to the cell surface. Methods in flow cytometry are discussed in Ormerod, M. G. (1994) Flow Cytometry, Oxford, New York NY.

The influence of MDDT on gene expression can be assessed using highly purified populations of cells transfected with sequences encoding MDDT and either CD64 or CD64-GFP. CD64 and CD64-GFP are expressed on the surface of transfected cells and bind to conserved regions of human immunoglobulin G (IgG). Transfected cells are efficiently separated from nontransfected cells using magnetic beads coated with either human IgG or antibody against CD64 (DYNAL, Inc., Lake Success NY). mRNA can be purified from the cells using methods well known by those of skill in the art. Expression of mRNA encoding MDDT and other genes of interest can be analyzed by northern analysis or microarray techniques.

XVI. Production of Antibodies

MDDT substantially purified using polyacrylamide gel electrophoresis (PAGE; see, e.g., Harrington, M.G. (1990) Methods Enzymol. 182:488-495), or other purification techniques, is used to immunize rabbits and to produce antibodies using standard protocols.

Alternatively, the MDDT amino acid sequence is analyzed using LASERGENE software (DNASTAR) to determine regions of high immunogenicity, and a corresponding peptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art. (See, e.g., Ausubel, 1995, supra, Chapter 11.)

Typically, peptides 15 residues in length are synthesized using an ABI 431A peptide synthesizer (PE Biosystems) using fmoc-chemistry and coupled to KLH (Sigma) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS) to increase immunogenicity. (See, e.g., Ausubel, supra.) Rabbits are immunized with the peptide-KLH complex in complete Freund's adjuvant. Resulting antisera are tested for antipeptide activity by, for example, binding the peptide to plastic, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG. Antisera with antipeptide activity are tested for anti-MDDT activity using protocols well known in the art, including ELISA, RIA, and immunoblotting.

XVII. Purification of Naturally Occurring MDDT Using Specific Antibodies

Naturally occurring or recombinant MDDT is substantially purified by immunoaffinity chromatography using antibodies specific for MDDT. An immunoaffinity column is constructed by covalently coupling anti-MDDT antibody to an activated chromatographic resin, such as CNBr-activated SEPHAROSE (Amersham Pharmacia Biotech). After the coupling, the resin is blocked and washed according to the manufacturer's instructions.

Media containing MDDT are passed over the immunoaffinity column, and the column is washed under conditions that allow the preferential absorbance of MDDT (e.g., high ionic strength buffers in the presence of detergent). The column is eluted under conditions that disrupt antibody/MDDT binding (e.g., a buffer of pH 2 to pH 3, or a high concentration of a chaotrope, such as urea or thiocyanate ion), and MDDT is collected.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention which are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

TABLE 1

SEQ ID NO:	Template ID	GI Number	Probability Score	Annotation
2	227709.3	g6468308	0	Novel human gene mapping to chomosome 1.
3	237703.2	g4103857	3.00E-73	fetal globin inducing factor (Mus musculus)
4	240091.1	g4038076	1.00E-141	Impact (Mus musculus)
5	243096.6	g2181494	1.00E-22	H.sapiens telomeric DNA sequence, clone 18QTEL072, read 18QTELOO072.seq.
7	405313.4	g6807590	8.00E-27	Novel human gene mapping to chomosome 1.
13	331642.1.j	g3879938	1.00E-28	Similarity to Human ADP/ATP carrier protein (SW:ADT1_HUMAN); cDNA EST EMBL:D71893 comes from this gene; cDNA EST EMBL:D74630 comes from this gene; cDNA EST yk394f1.3 comes from this gene; cDNA EST yk394f1.5 comes from this gene ()

TABLE 2

SEQ ID NO:	Template ID	Start	Stop	Frame	Pfam Hit	Pfam Description	E-value
1	222197.6	607	801	forward 1	zf-DHHC	DHHC zinc finger domain	2.40E-34
2	227709.3	76	432	forward 1	ArfGap	Putative GTP-ase activating protein for Arf	1.40E-52
3	237703.2	890	988	forward 2	ank	Ank repeat	8.80E-10
4	240091.1	366	905	forward 3	UPF0029	Uncharacterized protein family UPF0029	3.30E-12
5	243096.6	208	1056	forward 1	MMR_HSR1	GTPase of unknown function	6.40E-41
6	244366.6	611	805	forward 2	zf-DHHC	DHHC zinc finger domain	1.60E-32
7	405313.4	156	509	forward 3	ArfGap	Putative GTP-ase activating protein for Arf	1.30E-52
8	436857.2	400	1593	forward 1	Peptidase_M20	Peptidase family M20/M25/M40	8.90E-40
9	247285.1.j	372	461	forward 3	WW	WW domain	3.80E-04
10	254510.1.j	123	314	forward 3	KRAB	KRAB box	3.90E-24
11	284125.2.j	1069	1185	forward 1	WD40	WD domain, G-beta repeat	8.10E-10
12	331554.4.j	697	789	forward 1	PH	PH domain	4.80E-04
13	331642.1.j	446	757	forward 2	mito_carr	Mitochondrial carrier proteins	1.40E-05
13	33164.' 1.j	831	929	forward 3	mito_carr	Mitochondrial carrier proteins	4.20E-04
14	445594.2.j	155	283	forward 2	KRAB	KRAB box	4.20E-19

TABLE 3

SEQ ID NO:	Template ID	Start	Stop	Frame	Domain Type
1	222197.6	317	406	forward 2	SP
1	222197.6	901	984	forward 1	TM
2	227709.3	563	649	forward 2	SP
5	243096.6	3096	3182	forward 3	SP
6	244366.6	2801	2878	forward 2	TM
7	405313.4	2256	2333	forward 3	TM
7	405313.4	1503	1589	forward 3	TM

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
1	222197.6	3989355H1	1	122
1	222197.6	3989355R6	1	462
1	222197.6	g1189739	58	533
1	222197.6	g1123521	56	494
1	222197.6	3417884H2	105	341
1	222197.6	3398916H1	111	329
1	222197.6	696738H1	228	480
1	222197.6	3387328H1	248	542
1	222197.6	3387328F6	248	705
1	222197.6	640954H1	499	771
1	222197.6	640954R1	499	841
1	222197.6	2674395H1	544	647
1	222197.6	4871937H1	609	809
1	222197.6	6014949H1	680	954
1	222197.6	1310167H1	729	952
1	222197.6	1310167F6	729	1153
1	222197.6	3422058H1	733	986
1	222197.6	1429773H1	748	1016
1	222197.6	1429773F6	748	1211
1	222197.6	4725459H1	770	889
1	222197.6	2692245H1	774	1025
1	222197.6	2692245F6	774	1300
1	222197.6	2658283H1	818	1051
1	222197.6	4402233H1	847	1083
1	222197.6	673783H1	847	1089
1	222197.6	487422H1	871	1123
1	222197.6	3928678H1	898	1175
1	222197.6	2641613F6	1019	1494
1	222197.6	2641613H1	1019	1259
1	222197.6	2770396H1	1027	1273
1	222197.6	2599469H1	1040	1311
1	222197.6	486115H1	1181	1456
1	222197.6	1626615H1	1247	1456
1	222197.6	1626615F6	1247	1728
1	222197.6	383522H1	1260	1526
1	222197.6	3355867H1	1261	1532
1	222197.6	3617236H1	1286	1573
1	222197.6	3510978H1	1300	1567
1	222197.6	1568105H1	1325	1446
1	222197.6	1571377H1	1325	1550
1	222197.6	3806389H1	1368	1628
1	222197.6	g774888	1369	1729
1	222197.6	2995341H1	1382	1634
1	222197.6	5547807H1	1412	1611
1	222197.6	619375H1	1501	1738
1	222197.6	g1962367	1501	1997
1	222197.6	2695323H1	1517	1790
1	222197.6	3142880H1	1518	1792
1	222197.6	3805357H1	1518	1820
1	222197.6	1962884H1	1538	1808

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
1	222197.6	1807088F6	1555	2037
1	222197.6	1400677H1	1628	1894
1	222197.6	811652H1	1650	1954
1	222197.6	3048110H1	1665	1918
1	222197.6	3048110F6	1665	1992
1	222197.6	3048102H1	1665	1965
1	222197.6	5059358H1	1668	1965
1	222197.6	2150138H1	1671	1925
1	222197.6	039720H1	1711	1970
1	222197.6	2434795H1	3030	3132
1	222197.6	2130701H1	3039	3136
1	222197.6	2647674H1	1740	1841
1	222197.6	3448915H1	1810	2065
1	222197.6	4377215H1	1841	2043
1	222197.6	4317587H1	1841	1916
1	222197.6	1851036H1	1841	2033
1	222197.6	5108761H1	1861	2107
1	222197.6	2893265H1	1873	2139
1	222197.6	1568060H1	1878	2083
1	222197.6	1568004H1	1878	2097
1	222197.6	3531152H1	1889	2207
1	222197.6	2851378H1	1922	2260
1	222197.6	1931202H1	1922	2196
1	222197.6	2132091R6	1936	2208
1	222197.6	2132091H1	1936	2096
1	222197.6	5288578H1	1941	2067
1	222197.6	g2110737	1944	2226
1	222197.6	2207507H1	1996	2250
1	222197.6	2345452H1	1996	2256
1	222197.6	3414767H1	2025	2265
1	222197.6	1306171F6	2042	2378
1	222197.6	1306171H1	2042	2284
1	222197.6	4588038H1	2066	2349
1	222197.6	4587760H1	2066	2221
1	222197.6	1389765H1	2108	2367
1	222197.6	g1137612	2112	2425
1	222197.6	2663717H1	2146	2388
1	222197.6	3321090H1	2152	2435
1	222197.6	3840347H1	2151	2339
1	222197.6	2415559F6	2175	2618
1	222197.6	2415559H1	2175	2420
1	222197.6	3146224H1	2178	2430
1	222197.6	4201740H1	2178	2451
1	222197.6	3713261H1	2194	2446
1	222197.6	5900620H1	2195	2484
1	222197.6	1239238H1	2205	2358
1	222197.6	1965353R6	2235	2691
1	222197.6	1965353H1	2235	2500
1	222197.6	1471606H1	2273	2483
1	222197.6	3929839H1	2278	2578

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
1	222197.6	1521966H1	2278	2474
1	222197.6	1449385H1	2291	2533
1	222197.6	2381896H1	2307	2560
1	222197.6	2381895H1	2307	2559
1	222197.6	4643037H1	2326	2554
1	222197.6	3703243H1	2351	2650
1	222197.6	4295130H1	2350	2618
1	222197.6	4296184H1	2350	2589
1	222197.6	5841522H2	2396	2675
1	222197.6	3790395F6	2413	2974
1	222197.6	3811615H1	2413	2745
1	222197.6	2353349H1	2414	2513
1	222197.6	569817H1	2413	2660
1	222197.6	1621792H1	2413	2625
1	222197.6	520835H1	2417	2637
1	222197.6	g2161759	2433	2797
1	222197.6	1463438H1	2433	2622
1	222197.6	1621792T6	2437	3096
1	222197.6	3475011H1	2442	2682
1	222197.6	1969764H1	2447	2686
1	222197.6	4188958H1	2451	2774
1	222197.6	2134836H1	2458	2578
1	222197.6	4054390H1	2467	2749
1	222197.6	5185060H1	2468	2694
1	222197.6	4058390H1	2468	2580
1	222197.6	4024007H1	2469	2784
1	222197.6	5597388H1	2493	2771
1	222197.6	3934968H1	2512	2787
1	222197.6	2770396T6	2518	3095
1	222197.6	993964H1	2526	2698
1	222197.6	1807088T6	2531	3099
1	222197.6	2132091T6	2530	3101
1	222197.6	1965395T6	2532	3098
1	222197.6	1805709H1	2532	2781
1	222197.6	4466288H1	2537	2803
1	222197.6	3020435H1	2537	2821
1	222197.6	g2355832	2538	3035
1	222197.6	1672661H1	2554	2667
1	222197.6	1881147H1	2554	2807
1	222197.6	5098316H1	2573	2856
1	222197.6	1429773T6	2573	3089
1	222197.6	1626615T6	2584	3091
1	222197.6	1479854T6	2587	3117
1	222197.6	3935053H1	2598	2897
1	222197.6	3930918H1	2598	2915
1	222197.6	1654064H1	2608	2850
1	222197.6	2951301H1	2619	2908
1	222197.6	g4223642	2627	3028
1	222197.6	2752320H1	2628	2928
1	222197.6	g2161260	2634	3031

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
1	222197.6	1306171T6	2636	3099
1	222197.6	3387328T6	2640	3092
1	222197.6	811652T6	2639	3097
1	222197.6	1959841H1	2652	2915
1	222197.6	1959841T6	2652	3094
1	222197.6	1959841R6	2652	3110
1	222197.6	g4265714	2655	3143
1	222197.6	g4186863	2662	3139
1	222197.6	g3249761	2663	3146
1	222197.6	g4114970	2663	3136
1	222197.6	705080H1	2666	2908
1	222197.6	2045743H1	2673	2971
1	222197.6	2415559T6	2675	3097
1	222197.6	g4394362	2678	3067
1	222197.6	g2617967	2680	3140
1	222197.6	g1548565	2685	3028
1	222197.6	3790395T6	2684	3120
1	222197.6	g4393109	2684	3136
1	222197.6	1915761H1	2689	2948
1	222197.6	g2153774	2702	3136
1	222197.6	996350H1	2729	2980
1	222197.6	996350R1	2729	3028
1	222197.6	996350T1	2729	2992
1	222197.6	997484H1	2731	3034
1	222197.6	480176T6	2736	2990
1	222197.6	480176R6	2736	3028
1	222197.6	912187H1	2744	3042
1	222197.6	1313558H1	2744	3006
1	222197.6	g656198	2763	3141
1	222197.6	5057011H1	2784	3089
1	222197.6	2260766H1	2792	3062
1	222197.6	g3737413	2803	3139
1	222197.6	1686080H1	2810	3041
1	222197.6	g821623	2838	3148
1	222197.6	2641613T6	2833	3094
1	222197.6	2328044H1	2845	3113
1	222197.6	g4371777	2859	3141
1	222197.6	g1516072	2860	3144
1	222197.6	g2433045	2865	3091
1	222197.6	2648474H1	2866	3123
1	222197.6	2434653H1	2871	3069
1	222197.6	1878079H1	2881	3147
1	222197.6	g4109641	2896	3139
1	222197.6	2428514H1	2909	3098
1	222197.6	4146636H1	2909	3172
1	222197.6	4703838H1	2929	3139
1	222197.6	3125420H1	2934	3139
1	222197.6	5942277H1	2971	3137
2	227709.3	783646H1	1577	1867
2	227709.3	2314211H1	1585	1834

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	342368H1	1595	1832
2	227709.3	1833241R6	1595	2004
2	227709.3	1833241H1	1595	1853
2	227709.3	154117H1	1609	1752
2	227709.3	4938186H1	1617	1905
2	227709.3	4912871H1	1628	1918
2	227709.3	2243937H1	1653	1871
2	227709.3	876350H1	1530	1682
2	227709.3	4959782H1	1532	1785
2	227709.3	1839309H1	1663	1974
2	227709.3	1378304H1	1695	1940
2	227709.3	1660925H1	1696	1936
2	227709.3	1894917H1	1696	1912
2	227709.3	2394914H1	1716	1815
2	227709.3	3035674H1	1714	2034
2	227709.3	808578H1	1714	2019
2	227709.3	3035136H1	1714	2035
2	227709.3	808578R1	1714	2351
2	227709.3	3852902H1	1716	1984
2	227709.3	3122052H1	1729	2075
2	227709.3	6107590H1	1729	2046
2	227709.3	5925422H1	1735	2039
2	227709.3	4378762H1	1746	2059
2	227709.3	3781168H1	1763	1958
2	227709.3	2469542H1	1763	2006
2	227709.3	4585384H1	1765	2071
2	227709.3	3772159H1	1772	2087
2	227709.3	1436901F6	1787	2216
2	227709.3	1436902H1	1787	2081
2	227709.3	1436902F1	1787	2417
2	227709.3	732765H1	1796	2043
2	227709.3	531886H1	1796	2064
2	227709.3	732765R1	1796	2359
2	227709.3	323492H1	1796	2072
2	227709.3	1755142H1	1811	2072
2	227709.3	2088594H1	1817	2083
2	227709.3	1531927H1	1820	2036
2	227709.3	1281210H1	1828	1963
2	227709.3	618185H1	1834	2133
2	227709.3	072066H1	1833	2072
2	227709.3	920524H1	1837	2173
2	227709.3	g2030053	1840	2258
2	227709.3	g681548	1845	2248
2	227709.3	g1190789	1847	2173
2	227709.3	3052574H1	1853	2158
2	227709.3	4546684H1	1856	1963
2	227709.3	g1846206	1855	2184
2	227709.3	1231442H1	1856	2170
2	227709.3	4546692H1	1856	1961
2	227709.3	1231220H1	1856	2108

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	5693680H1	1861	2153
2	227709.3	2040451H1	1861	2194
2	227709.3	5781387H1	1861	2129
2	227709.3	3506145H1	1861	2181
2	227709.3	3479633H1	1859	2214
2	227709.3	2196589H1	1861	2123
2	227709.3	3872090H1	1867	2086
2	227709.3	1210943R1	1867	2217
2	227709.3	072676H1	1867	2104
2	227709.3	3120279H1	1867	2163
2	227709.3	1210943H1	1867	2127
2	227709.3	5877032H1	1869	2133
2	227709.3	5469466H1	1875	2177
2	227709.3	030331H1	1877	2151
2	227709.3	g1970048	1878	2203
2	227709.3	2257012H1	1883	2139
2	227709.3	705952H1	1891	2209
2	227709.3	2564249H1	1891	2202
2	227709.3	5712256H1	1902	2222
2	227709.3	3798289H1	1903	2214
2	227709.3	456952H1	1903	2171
2	227709.3	693406H1	1910	2219
2	227709.3	2403540H1	1915	2217
2	227709.3	6095157H1	1918	2217
2	227709.3	4257073H1	1928	2230
2	227709.3	074494H1	1936	2178
2	227709.3	073044H1	1936	2246
2	227709.3	073608H1	1936	2217
2	227709.3	5882532H1	1937	2217
2	227709.3	073991H1	1936	2227
2	227709.3	073890H1	1936	2119
2	227709.3	073335H1	1936	2162
2	227709.3	5882935H1	1938	2217
2	227709.3	5883716H1	1938	2217
2	227709.3	5881208H1	1939	2217
2	227709.3	5888519H1	1939	2212
2	227709.3	5890218H1	1939	2212
2	227709.3	4783188H1	1938	2225
2	227709.3	2317709H1	1941	2222
2	227709.3	1876721H1	1952	2217
2	227709.3	2469335H1	1954	2223
2	227709.3	734056H1	1953	2076
2	227709.3	2938267H1	1957	2217
2	227709.3	3166569H1	1957	2217
2	227709.3	4591368H1	1966	2227
2	227709.3	2397855H1	1966	2240
2	227709.3	6105204H1	1965	2217
2	227709.3	874849H1	1968	2217
2	227709.3	4458852H1	1967	2217
2	227709.3	874849R1	1968	2621

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	1896188H1	1989	2217
2	227709.3	2470872T6	1988	2644
2	227709.3	1831904H1	1994	2217
2	227709.3	2433378H1	2003	2173
2	227709.3	4893472H1	2003	2337
2	227709.3	2395264H1	2016	2217
2	227709.3	4203434H1	2018	2350
2	227709.3	4886606H1	2025	2329
2	227709.3	4886606F6	2025	2094
2	227709.3	504991H1	2036	2217
2	227709.3	3123928H1	2035	2317
2	227709.3	372978R6	2539	2691
2	227709.3	2356168H1	2576	2698
2	227709.3	213804H1	2632	2687
2	227709.3	3347292H1	461	695
2	227709.3	5013953H1	518	793
2	227709.3	4327103H1	527	781
2	227709.3	3295045H1	569	811
2	227709.3	4012389H1	621	883
2	227709.3	1377181F1	637	1043
2	227709.3	1377181H1	637	880
2	227709.3	5883912H1	663	856
2	227709.3	5886832H1	663	927
2	227709.3	5881250H1	664	943
2	227709.3	377379H1	687	946
2	227709.3	2098327H1	697	942
2	227709.3	3511206H1	728	873
2	227709.3	4031927H1	736	993
2	227709.3	2448606H1	738	977
2	227709.3	2473593T6	743	1332
2	227709.3	2444327H1	744	975
2	227709.3	388684H1	774	1038
2	227709.3	924593R1	778	1154
2	227709.3	924593H1	778	1044
2	227709.3	2707537T6	791	1332
2	227709.3	1842323T6	803	1331
2	227709.3	1386726H1	813	1109
2	227709.3	1436357H1	816	1076
2	227709.3	1436357F1	816	1379
2	227709.3	1842323H1	818	1009
2	227709.3	1842323R6	818	1347
2	227709.3	2757452H1	862	1137
2	227709.3	338917H1	874	1099
2	227709.3	g1382744	934	1319
2	227709.3	g2880866	952	1325
2	227709.3	5289932H1	954	1212
2	227709.3	736987R6	967	1219
2	227709.3	g2955000	966	1369
2	227709.3	736987H1	967	1187
2	227709.3	4792654H1	971	1250

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	270359H1	989	1337
2	227709.3	2532882H1	985	1263
2	227709.3	3246655H1	990	1254
2	227709.3	g2106694	1001	1373
2	227709.3	6210707H1	1069	1397
2	227709.3	340933H1	1086	1262
2	227709.3	4638914H1	1094	1344
2	227709.3	6211794H1	1106	1397
2	227709.3	1220278H1	1148	1408
2	227709.3	4762916H1	1193	1486
2	227709.3	6208765H1	1202	1504
2	227709.3	3167466H1	1208	1497
2	227709.3	4541633H1	1218	1466
2	227709.3	1389793H1	1	242
2	227709.3	1221361H1	146	324
2	227709.3	6210207H1	1217	1524
2	227709.3	6208587H1	1217	1448
2	227709.3	432063H1	149	454
2	227709.3	2197686H1	1241	1389
2	227709.3	3871923H1	142	426
2	227709.3	2473593H1	205	438
2	227709.3	014086H1	1244	1539
2	227709.3	g1809628	1244	1573
2	227709.3	g570725	1274	1510
2	227709.3	013903H1	1285	1535
2	227709.3	3159468H1	1288	1590
2	227709.3	5219130H1	1303	1575
2	227709.3	2473593F6	205	346
2	227709.3	862771H1	1304	1579
2	227709.3	3368510H1	1323	1609
2	227709.3	2473008H1	1340	1591
2	227709.3	3633133H1	1353	1666
2	227709.3	2470872F6	267	453
2	227709.3	2472485H1	1367	1615
2	227709.3	3940725H1	1371	1561
2	227709.3	1839612H1	1376	1668
2	227709.3	1839635H1	1376	1702
2	227709.3	2440859H1	1388	1640
2	227709.3	2560406H1	1389	1677
2	227709.3	g776447	1411	1595
2	227709.3	g892952	1432	1807
2	227709.3	4753851H1	1460	1734
2	227709.3	855862R1	1460	2074
2	227709.3	855862H1	1460	1682
2	227709.3	5436709H1	1470	1708
2	227709.3	3872292H1	1474	1684
2	227709.3	2470872H1	267	518
2	227709.3	4738304H2	363	616
2	227709.3	634613H1	376	613
2	227709.3	4890310H1	1474	1746

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	964305H1	396	658
2	227709.3	g1303518	1483	2011
2	227709.3	4081986H1	1497	1695
2	227709.3	2731973H1	1501	1745
2	227709.3	964305R1	396	999
2	227709.3	4984432H1	398	674
2	227709.3	4323306H1	407	691
2	227709.3	656683H1	431	693
2	227709.3	656689H1	431	723
2	227709.3	g774837	1512	1815
2	227709.3	g573087	1512	1840
2	227709.3	3038419H1	1514	1800
2	227709.3	2830914H1	1515	1788
2	227709.3	5206976H2	1530	1806
2	227709.3	073335T6	2039	2650
2	227709.3	g3173537	2054	2217
2	227709.3	862087T1	2058	2649
2	227709.3	862087H1	2058	2344
2	227709.3	g1955564	2058	2421
2	227709.3	2469827T6	2058	2648
2	227709.3	2395978H1	2063	2318
2	227709.3	1833241T6	2074	2648
2	227709.3	1226609H1	2075	2353
2	227709.3	1436901T6	2075	2641
2	227709.3	2827153H1	2083	2449
2	227709.3	1878980H1	2083	2374
2	227709.3	2473240T6	2083	2643
2	227709.3	2431389H1	2087	2329
2	227709.3	2400824H1	2087	2358
2	227709.3	1879688F6	2095	2530
2	227709.3	1879688H1	2095	2389
2	227709.3	1879688T6	2096	2653
2	227709.3	2195425H1	2098	2400
2	227709.3	g1962618	2102	2694
2	227709.3	2325847H1	2112	2374
2	227709.3	4468886H1	2112	2414
2	227709.3	3940725T6	2118	2654
2	227709.3	2917539H1	2119	2418
2	227709.3	1347063H1	2130	2371
2	227709.3	554568H1	2137	2387
2	227709.3	323431H1	2148	2442
2	227709.3	450205H1	2150	2378
2	227709.3	1448863H1	2150	2423
2	227709.3	2472138T6	2186	2646
2	227709.3	736987T6	2206	2648
2	227709.3	g3932020	2237	2687
2	227709.3	1676401H1	2247	2475
2	227709.3	406441H1	2259	2522
2	227709.3	334657H1	2259	2512
2	227709.3	5888253H1	2261	2496

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	3102430H1	2261	2530
2	227709.3	6106355H1	2261	2524
2	227709.3	g3182016	2261	2693
2	227709.3	604792H1	2261	2486
2	227709.3	2428425H1	2261	2433
2	227709.3	5883372H1	2261	2444
2	227709.3	g4264131	2262	2702
2	227709.3	633186H1	2262	2544
2	227709.3	3993266H1	2261	2540
2	227709.3	2017401H1	2263	2427
2	227709.3	2424057H1	2268	2536
2	227709.3	404680H1	2273	2496
2	227709.3	3153874H1	2288	2593
2	227709.3	g4291477	2288	2694
2	227709.3	4061504H1	2288	2546
2	227709.3	g2968506	2289	2694
2	227709.3	3123730H1	2289	2606
2	227709.3	3123955H1	2289	2589
2	227709.3	g4115080	2293	2694
2	227709.3	308633H1	2295	2534
2	227709.3	2397634H1	2297	2495
2	227709.3	g4332177	2302	2694
2	227709.3	g4267824	2303	2694
2	227709.3	308633F1	2303	2687
2	227709.3	308633R1	2303	2687
2	227709.3	2371681H1	2309	2551
2	227709.3	2421774H1	2309	2547
2	227709.3	g1810042	2311	2687
2	227709.3	g3958397	2313	2677
2	227709.3	4695290H1	2316	2582
2	227709.3	2756565H1	2320	2630
2	227709.3	g612404	2323	2687
2	227709.3	2445854H1	2327	2587
2	227709.3	449703H1	2332	2458
2	227709.3	2877181H1	2334	2622
2	227709.3	1211271R1	2334	2687
2	227709.3	1211271T1	2334	2649
2	227709.3	1211271H1	2334	2608
2	227709.3	g616409	2341	2660
2	227709.3	2017458H1	2340	2619
2	227709.3	1534375H1	2342	2572
2	227709.3	3145020H1	2344	2683
2	227709.3	1539734H1	2358	2600
2	227709.3	3786174H1	2363	2661
2	227709.3	1454741F1	2367	2687
2	227709.3	1454741H1	2367	2632
2	227709.3	2717951H1	2372	2548
2	227709.3	1359816H1	2374	2618
2	227709.3	1359816F1	2374	2694
2	227709.3	g564645	2385	2694

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
2	227709.3	g1154316	2388	2698
2	227709.3	g891561	2393	2707
2	227709.3	2473312H1	2393	2648
2	227709.3	g2992779	2394	2643
2	227709.3	1218580T6	2395	2649
2	227709.3	1218580T1	2395	2649
2	227709.3	g824343	2396	2719
2	227709.3	1218580R6	2395	2691
2	227709.3	1218573H1	2395	2641
2	227709.3	286961H1	2403	2690
2	227709.3	4367190H1	2408	2684
2	227709.3	2399122H1	2420	2672
2	227709.3	5882896H1	2423	2687
2	227709.3	g967282	2422	2700
2	227709.3	5883904H1	2423	2687
2	227709.3	5883908H1	2423	2687
2	227709.3	5882375H1	2424	2567
2	227709.3	g1191305	2429	2710
2	227709.3	794290H1	2437	2664
2	227709.3	520742H1	2442	2678
2	227709.3	g646174	2443	2687
2	227709.3	1538631H1	2443	2651
2	227709.3	1722285H1	2444	2677
2	227709.3	g1202716	2469	2701
2	227709.3	862903T1	2482	2648
2	227709.3	095638H1	2484	2694
2	227709.3	862903R1	2484	2694
2	227709.3	3124846H1	2490	2692
2	227709.3	5906470H1	2497	2691
2	227709.3	2535162H1	2525	2651
2	227709.3	4460277H1	2533	2694
2	227709.3	372978T6	2539	2649
2	227709.3	372978H1	2539	2686
3	237703.2	g1963754	1	374
3	237703.2	g1137733	95	407
3	237703.2	g843567	124	414
3	237703.2	3070350H1	189	477
3	237703.2	3070350F6	189	709
3	237703.2	1439542H1	413	686
3	237703.2	3203352H1	437	711
3	237703.2	g2013304	598	954
3	237703.2	3799002H1	701	1010
3	237703.2	044160T6	958	1453
3	237703.2	824258H1	1020	1254
3	237703.2	g189439	1031	1472
3	237703.2	3491432H1	1052	1315
3	237703.2	2601554F6	1059	1602
3	237703.2	2601554H1	1060	1336
3	237703.2	4617816H1	1073	1337
3	237703.2	4058186H1	1111	1197

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
3	237703.2	5554248H1	1145	1355
3	237703.2	5554148H1	1145	1386
3	237703.2	g3594366	1199	1611
3	237703.2	5122547H1	1278	1516
3	237703.2	2278690H1	1379	1654
3	237703.2	2278690R6	1379	1879
3	237703.2	5564683H1	1402	1658
3	237703.2	950519H1	1436	1678
3	237703.2	950519R6	1436	1723
3	237703.2	g1123529	1437	1805
3	237703.2	5260937H1	1514	1744
3	237703.2	3386649H1	1527	1744
3	237703.2	5272970H1	1532	1780
3	237703.2	3808872H1	1533	1838
3	237703.2	950519T6	1554	2019
3	237703.2	319252H1	1592	1986
3	237703.2	4411457H1	1669	1947
3	237703.2	g1933240	1669	2147
3	237703.2	2601554T6	1698	2313
3	237703.2	824257T6	1699	2313
3	237703.2	2859138T6	1776	2312
3	237703.2	1872409F6	1788	2150
3	237703.2	1872409H1	1788	2062
3	237703.2	1572418H1	1798	1997
3	237703.2	1872409T6	1805	2312
3	237703.2	530381H1	1810	1963
3	237703.2	5583255H1	1811	2075
3	237703.2	126942H1	1816	2025
3	237703.2	2278690T6	1829	2311
3	237703.2	1211984H1	1832	2066
3	237703.2	2703527H1	1832	2106
3	237703.2	3253906H1	1885	2159
3	237703.2	g3934221	1896	2349
3	237703.2	1620273H1	1897	2117
3	237703.2	g2819399	1908	2351
3	237703.2	g3895924	1936	2349
3	237703.2	g2881190	1962	2270
3	237703.2	040587H1	1971	2158
3	237703.2	g3147053	1973	2349
3	237703.2	g843522	2009	2349
3	237703.2	g1844904	2023	2349
3	237703.2	g2881790	2024	2349
3	237703.2	g2820075	2030	2349
3	237703.2	g2237723	2051	2350
3	237703.2	238512H1	2085	2313
3	237703.2	292954H1	2182	2320
3	237703.2	g2013921	2238	2522
3	237703.2	g1980268	2386	2742
4	240091.1	2898155H1	1	289
4	240091.1	2434264H1	3	215

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
4	240091.1	5075278H1	3	127
4	240091.1	2647594H1	1720	1770
4	240091.1	4785026H1	1803	2073
4	240091.1	4785001H1	1803	2069
4	240091.1	3600323H1	1258	1557
4	240091.1	2448218F6	1267	1485
4	240091.1	2448218H1	1267	1508
4	240091.1	1391961T6	1267	1730
4	240091.1	083349H1	1309	1464
4	240091.1	070643H1	1309	1543
4	240091.1	3712971T6	1317	1744
4	240091.1	3399693H1	1384	1606
4	240091.1	g3000643	1400	1562
4	240091.1	g3659260	1446	1772
4	240091.1	3491102H1	1447	1559
4	240091.1	2286846H1	1496	1696
4	240091.1	4575130H1	1513	1756
4	240091.1	2584803H1	1539	1770
4	240091.1	2584803F6	1539	1770
4	240091.1	4399541H1	1590	1833
4	240091.1	489490H1	1599	1844
4	240091.1	5541073H1	1605	1804
4	240091.1	797486H1	1609	1772
4	240091.1	2435453H1	3	231
4	240091.1	2434264R6	3	491
4	240091.1	527914H1	4	275
4	240091.1	4382936H1	10	241
4	240091.1	4210908H1	20	292
4	240091.1	3615238H1	47	340
4	240091.1	3615238F6	47	528
4	240091.1	2733107H1	54	275
4	240091.1	494380H1	62	307
4	240091.1	1391961F6	66	475
4	240091.1	1391961H1	66	318
4	240091.1	580112H1	359	558
4	240091.1	1232706F6	389	842
4	240091.1	1232706H1	389	629
4	240091.1	3487133H1	432	698
4	240091.1	g4244249	511	981
4	240091.1	447619H1	580	799
4	240091.1	5782214H1	670	964
4	240091.1	4913546F6	830	1249
4	240091.1	4913546H1	830	1108
4	240091.1	3892111H1	834	1130
4	240091.1	4742244H1	836	1102
4	240091.1	g1484624	867	1316
4	240091.1	2376485F6	901	1205
4	240091.1	2376485H1	901	1124
4	240091.1	2376485T6	902	1167
4	240091.1	1849607H1	907	1198

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
4	240091.1	4030992H1	952	1198
4	240091.1	2764886H1	975	1201
4	240091.1	4665206H1	1138	1401
4	240091.1	1232706T6	1176	1729
4	240091.1	2434264T6	1210	1746
4	240091.1	4424318H1	1223	1459
5	243096.6	g3228919	1006	1482
5	243096.6	g3597815	1016	1479
5	243096.6	g3933703	1018	1479
5	243096.6	g4372582	1022	1487
5	243096.6	4103071H1	1030	1104
5	243096.6	3478526H1	1032	1183
5	243096.6	g4457447	1045	1481
5	243096.6	g4222324	1075	1480
5	243096.6	631209R6	1076	1478
5	243096.6	g3118595	1075	1478
5	243096.6	g2577165	1076	1480
5	243096.6	g2185952	1078	1493
5	243096.6	g2063697	1079	1482
5	243096.6	222052H1	1079	1215
5	243096.6	222052F1	1078	1479
5	243096.6	222052R1	1078	1479
5	243096.6	g3737532	1083	1509
5	243096.6	1849724T6	1095	1440
5	243096.6	g4110131	1115	1481
5	243096.6	631209T6	1116	1438
5	243096.6	2446727T6	1119	1438
5	243096.6	g3155321	1129	1475
5	243096.6	g3178176	1148	1494
5	243096.6	948177H1	1149	1428
5	243096.6	948177R1	1149	1488
5	243096.6	1942176H1	1163	1441
5	243096.6	1942176R6	1163	1418
5	243096.6	1942168H1	1163	1440
5	243096.6	5884138H1	1164	1433
5	243096.6	1418514T6	1209	1430
5	243096.6	1418514H1	1216	1431
5	243096.6	1418364H1	1216	1468
5	243096.6	1418514F6	1216	1479
5	243096.6	632343H1	1218	1458
5	243096.6	g4107711	1220	1526
5	243096.6	g4457962	1227	1483
5	243096.6	g2837785	1228	1479
5	243096.6	g819991	1238	1496
5	243096.6	g564440	1237	1488
5	243096.6	g816379	1251	1540
5	243096.6	g885380	1252	1488
5	243096.6	g768804	1261	1481
5	243096.6	6093263H1	1263	1492
5	243096.6	g645318	1286	1488

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
5	243096.6	g566867	1292	1526
5	243096.6	g816311	1302	1679
5	243096.6	g671079	1296	1488
5	243096.6	g2219072	1297	1472
5	243096.6	g2539665	1300	1479
5	243096.6	g670466	1300	1526
5	243096.6	2474482H1	1313	1542
5	243096.6	g4328047	1314	1487
5	243096.6	g2205935	1335	1454
5	243096.6	g832021	1367	1678
5	243096.6	g2205936	1375	1472
5	243096.6	g2789365	1393	1479
5	243096.6	g873007	1418	1540
5	243096.6	g900098	1419	1488
5	243096.6	g567639	1424	1667
5	243096.6	1918362H1	1473	1731
5	243096.6	4727611H1	1576	1854
5	243096.6	g822245	1648	1976
5	243096.6	g812869	1651	2012
5	243096.6	g830918	1654	2012
5	243096.6	1414612H1	1662	1890
5	243096.6	4761250H1	1662	1939
5	243096.6	g678372	1665	1972
5	243096.6	g561207	1665	1955
5	243096.6	g2002379	1665	2002
5	243096.6	g709471	1665	1866
5	243096.6	4761242H1	1664	1939
5	243096.6	g518391	1678	1933
5	243096.6	4595686H1	1707	1981
5	243096.6	1511493H1	1792	1996
5	243096.6	1511493F6	1792	2244
5	243096.6	1512376H1	1792	2009
5	243096.6	g2003356	1922	2087
5	243096.6	4697285H1	2203	2446
5	243096.6	4941432H1	2381	2673
5	243096.6	1230891H1	2413	2508
5	243096.6	1522037H1	2421	2625
5	243096.6	3749969H1	2502	2799
5	243096.6	2125142H1	2527	2795
5	243096.6	2125142F6	2527	2841
5	243096.6	121562H1	2620	2806
5	243096.6	856668H1	2745	2933
5	243096.6	5882521H1	2758	3030
5	243096.6	5888582H1	2759	2976
5	243096.6	5882569H1	2760	3030
5	243096.6	g775350	2793	3137
5	243096.6	g705857	2790	3138
5	243096.6	g2002380	2803	3138
5	243096.6	5927949H1	2845	3140
5	243096.6	1511493T6	2883	3500

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
5	243096.6	1335311H1	2951	3205
5	243096.6	1613745H1	2965	3179
5	243096.6	3472823H1	3005	3245
5	243096.6	g570224	3048	3318
5	243096.6	g4095588	3134	3545
5	243096.6	g831152	3143	3366
5	243096.6	g4286632	3205	3476
5	243096.6	5907720H1	3208	3501
5	243096.6	g4187457	3286	3557
5	243096.6	g3842315	3295	3465
5	243096.6	g4005713	3303	3465
5	243096.6	g4006389	3305	3465
5	243096.6	g4006377	3305	3559
5	243096.6	g4006150	3305	3537
5	243096.6	g4006070	3315	3542
5	243096.6	g4187003	3315	3554
5	243096.6	g4188554	3315	3537
5	243096.6	g4006771	3315	3537
5	243096.6	g4072007	3316	3542
5	243096.6	g4017934	3316	3537
5	243096.6	g4150328	3316	3465
5	243096.6	g4005644	3316	3537
5	243096.6	5840086H1	3345	3553
5	243096.6	5289394H1	3472	3737
5	243096.6	g710217	3508	3789
5	243096.6	g694295	3619	3781
5	243096.6	g2206232	3623	3794
5	243096.6	g2206104	3656	3795
5	243096.6	2897215H1	1	249
5	243096.6	3541808H1	181	397
5	243096.6	2352032H1	32	249
5	243096.6	2446727F6	44	104
5	243096.6	3123367H1	44	356
5	243096.6	4385825H1	181	379
5	243096.6	2446727H1	44	308
5	243096.6	2905666H1	45	326
5	243096.6	2767616H1	46	308
5	243096.6	4521271H1	214	473
5	243096.6	1725750H1	47	209
5	243096.6	g1965606	235	621
5	243096.6	3117919H1	47	328
5	243096.6	2762827H1	49	309
5	243096.6	5395762H1	245	510
5	243096.6	5585677H1	251	484
5	243096.6	3416289H1	253	507
5	243096.6	5407275H1	257	511
5	243096.6	5407149H1	257	520
5	243096.6	3452689H1	49	240
5	243096.6	4819033H1	292	515
5	243096.6	483458H1	50	302

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
5	243096.6	1941753H1	291	537
5	243096.6	485741H1	50	296
5	243096.6	g766717	307	480
5	243096.6	2215901H1	51	147
5	243096.6	2990593H1	73	383
5	243096.6	4017755H1	76	378
5	243096.6	2558416H1	81	363
5	243096.6	870507R1	82	682
5	243096.6	870507H1	82	339
5	243096.6	3692401H1	82	285
5	243096.6	2387055H1	83	336
5	243096.6	1641267H1	329	555
5	243096.6	2483682H1	84	331
5	243096.6	4977750H1	89	382
5	243096.6	g4244257	343	817
5	243096.6	3165660H1	89	373
5	243096.6	4043243H1	89	406
5	243096.6	3500940H1	347	625
5	243096.6	3580985H1	91	415
5	243096.6	1518953F6	93	405
5	243096.6	g672832	92	414
5	243096.6	g574622	92	418
5	243096.6	2206923H1	355	624
5	243096.6	2681788H1	92	287
5	243096.6	g672843	92	444
5	243096.6	790680R1	365	938
5	243096.6	790680H1	365	584
5	243096.6	3500449H1	386	701
5	243096.6	1518953H1	92	280
5	243096.6	3510335H1	92	396
5	243096.6	4401867H1	393	653
5	243096.6	1624276H1	395	583
5	243096.6	3099469H1	92	415
5	243096.6	g873107	93	484
5	243096.6	g874944	93	492
5	243096.6	2201243H1	411	667
5	243096.6	4907323H2	97	377
5	243096.6	2215590H1	421	665
5	243096.6	1647105H1	102	323
5	243096.6	3337242H1	105	332
5	243096.6	3328567H1	421	709
5	243096.6	5165830H1	113	391
5	243096.6	1919378R6	432	865
5	243096.6	2078775H1	114	391
5	243096.6	1919378H1	432	700
5	243096.6	1798353H1	115	371
5	243096.6	5109893H1	447	675
5	243096.6	3581083H1	116	378
5	243096.6	2202470H1	456	711
5	243096.6	1642210H1	461	676

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
5	243096.6	2828817H1	122	399
5	243096.6	1642206H1	461	676
5	243096.6	g669309	136	449
5	243096.6	4932616H1	463	618
5	243096.6	g571269	136	492
5	243096.6	1753890H1	464	690
5	243096.6	756217H1	464	706
5	243096.6	3030389H1	464	764
5	243096.6	g677690	136	462
5	243096.6	1754027H1	464	704
5	243096.6	3893562H1	139	449
5	243096.6	g885379	139	482
5	243096.6	026083H1	509	692
5	243096.6	2758492H1	143	413
5	243096.6	836506R1	513	1092
5	243096.6	5294950H1	147	395
5	243096.6	836506H1	513	759
5	243096.6	1520256H1	517	685
5	243096.6	173031H1	157	390
5	243096.6	5197820H2	161	419
5	243096.6	g2240993	522	940
5	243096.6	g766746	161	395
5	243096.6	3728525H1	572	854
5	243096.6	g677072	160	502
5	243096.6	2828115T6	574	950
5	243096.6	g2816446	612	874
5	243096.6	4536339H1	624	877
5	243096.6	2670757H1	629	868
5	243096.6	4994228H1	697	1004
5	243096.6	793596H1	697	946
5	243096.6	g2058963	697	941
5	243096.6	1560602H1	719	948
5	243096.6	1535660H1	719	898
5	243096.6	g2058866	730	936
5	243096.6	2316449H1	764	1045
5	243096.6	686656H1	771	1028
5	243096.6	3728525T1	783	1432
5	243096.6	3659224H2	789	1073
5	243096.6	639411H1	798	1050
5	243096.6	5332257H1	820	1063
5	243096.6	2084254H1	837	1136
5	243096.6	2652025T6	859	1424
5	243096.6	961879T6	859	1437
5	243096.6	5088178T6	858	1465
5	243096.6	1849724F6	860	1441
5	243096.6	1849724H1	860	1135
5	243096.6	5395762T1	884	1440
5	243096.6	1919378T6	880	1452
5	243096.6	2663785H1	891	1144
5	243096.6	3625012H1	904	1051

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
5	243096.6	2683022H1	931	1212
5	243096.6	3499439H1	933	1219
5	243096.6	4005888H1	935	1211
5	243096.6	1682469T7	939	1485
5	243096.6	2351530H1	941	1130
5	243096.6	g2063947	951	1207
5	243096.6	1668866H1	960	1201
5	243096.6	1667633H1	960	1220
5	243096.6	g2358923	963	1060
5	243096.6	3085780H1	971	1082
5	243096.6	g814507	161	443
5	243096.6	g830479	161	470
5	243096.6	g816410	162	519
6	244366.6	1889554H1	493	750
6	244366.6	1889554F6	493	939
6	244366.6	4298324H1	1	255
6	244366.6	853003H1	12	225
6	244366.6	g2178494	517	912
6	244366.6	853003R6	26	483
6	244366.6	3327565H1	555	787
6	244366.6	2263295H1	278	532
6	244366.6	5401026H1	604	816
6	244366.6	1285225H1	606	862
6	244366.6	2674162H1	661	904
6	244366.6	3101288H1	295	585
6	244366.6	3295139H1	815	1056
6	244366.6	6002940H1	886	1170
6	244366.6	3101288F6	295	694
6	244366.6	6002740H1	904	1170
6	244366.6	3246058F6	941	1282
6	244366.6	3246058H1	941	1192
6	244366.6	3887233H1	959	1240
6	244366.6	2431320H1	972	1192
6	244366.6	1513444H1	978	1189
6	244366.6	2813740H1	1071	1363
6	244366.6	2815664H1	1071	1274
6	244366.6	2813707H1	1071	1359
6	244366.6	3492628H1	1138	1414
6	244366.6	2183893H1	1190	1450
6	244366.6	5641164H1	1238	1485
6	244366.6	580082H1	1239	1487
6	244366.6	3155135H1	1254	1487
6	244366.6	3075416H1	1282	1565
6	244366.6	3559024H1	1351	1639
6	244366.6	3451987H1	1525	1785
6	244366.6	4378692H1	1597	1817
6	244366.6	g2162961	1742	2237
6	244366.6	3890528H1	1764	1919
6	244366.6	5017346H1	3006	3272
6	244366.6	1690531H1	2938	3105

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
6	244366.6	5890083H1	3007	3133
6	244366.6	4614333H1	3009	3145
6	244366.6	2783117H1	2979	3239
6	244366.6	3037603H1	3012	3292
6	244366.6	032789H1	3013	3234
6	244366.6	5262767H2	3029	3291
6	244366.6	2626903H1	3031	3283
6	244366.6	1403785H1	3034	3328
6	244366.6	3321234H1	2980	3254
6	244366.6	003727H1	3043	3400
6	244366.6	003176H1	3043	3543
6	244366.6	003684H1	3043	3424
6	244366.6	003185H1	3043	3550
6	244366.6	003182H1	3043	3493
6	244366.6	003701H1	3043	3471
6	244366.6	003615H1	3043	3429
6	244366.6	003188H1	3043	3483
6	244366.6	003127H1	3043	3453
6	244366.6	003465H1	3043	3433
6	244366.6	003422H1	3043	3380
6	244366.6	5290617H1	3002	3300
6	244366.6	003521H1	3043	3549
6	244366.6	003294H1	3043	3411
6	244366.6	003642H1	3043	3400
6	244366.6	003646H1	3043	3405
6	244366.6	003660H1	3043	3392
6	244366.6	5138512H1	3079	3394
6	244366.6	094605H1	3081	3258
6	244366.6	1726602H1	3114	3335
6	244366.6	1623474T6	3120	3724
6	244366.6	2381350H1	3122	3380
6	244366.6	768275H1	3130	3388
6	244366.6	3495251H1	3143	3350
6	244366.6	3705742H1	3161	3533
6	244366.6	4744563H1	3175	3471
6	244366.6	3101288T6	3197	3715
6	244366.6	5561382H1	3201	3503
6	244366.6	074734H1	3202	3442
6	244366.6	073329H1	3202	3403
6	244366.6	1975441T6	3222	3706
6	244366.6	3477744H1	3244	3416
6	244366.6	2112529T6	3250	3723
6	244366.6	g3933445	3274	3756
6	244366.6	4861989H1	3274	3567
6	244366.6	1737024F6	3292	3734
6	244366.6	g2584374	3285	3757
6	244366.6	1735490H1	3292	3561
6	244366.6	1737024H1	3292	3554
6	244366.6	393035H1	3303	3590
6	244366.6	2158031F6	3307	3758

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
6	244366.6	g4438605	3308	3758
6	244366.6	2554055H1	3325	3624
6	244366.6	g2934389	3326	3756
6	244366.6	g4391414	3341	3756
6	244366.6	4460645H1	3349	3601
6	244366.6	g2208607	3359	3756
6	244366.6	g2318342	3366	3756
6	244366.6	g1062585	3370	3743
6	244366.6	g4081771	3372	3761
6	244366.6	g1925212	3381	3757
6	244366.6	4726758H1	3390	3662
6	244366.6	g2410378	3392	3763
6	244366.6	867419H1	3398	3679
6	244366.6	g616070	3402	3764
6	244366.6	g2163418	3405	3756
6	244366.6	g2555602	3413	3760
6	244366.6	g561365	3428	3756
6	244366.6	1889554T6	3433	3725
6	244366.6	g2336915	3444	3756
6	244366.6	g616115	3445	3756
6	244366.6	g4435130	3469	3756
6	244366.6	g4525507	3502	3757
6	244366.6	2158031H1	3523	3756
6	244366.6	g2401624	3528	3765
6	244366.6	g4268526	3559	3756
6	244366.6	2009370H1	3666	3756
6	244366.6	218073H1	3682	3756
6	244366.6	2350763H1	3699	3760
6	244366.6	1647483H1	2520	2769
6	244366.6	2432662H1	2526	2757
6	244366.6	901941R1	2538	3096
6	244366.6	901941H1	2538	2895
6	244366.6	901981H1	2538	2858
6	244366.6	2052263H1	2545	2857
6	244366.6	3483573H1	2553	2851
6	244366.6	3565807H1	2558	2822
6	244366.6	g2278841	2560	2917
6	244366.6	g2178439	2563	2917
6	244366.6	g2153824	2565	2917
6	244366.6	g1329145	2568	2874
6	244366.6	2198515H1	2647	2908
6	244366.6	2200581H1	2647	2725
6	244366.6	g1548506	2680	3207
6	244366.6	2321185H1	2694	2917
6	244366.6	3244071T6	2568	2798
6	244366.6	2936492H1	2700	2917
6	244366.6	600642H1	2586	2891
6	244366.6	g1124072	2711	2850
6	244366.6	g1833465	2712	2856
6	244366.6	g4327019	2736	2851

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
6	244366.6	3165778H1	2588	2925
6	244366.6	g2139164	2594	2835
6	244366.6	633565H1	2753	2917
6	244366.6	1520269F6	2766	3147
6	244366.6	1520026H1	2766	2917
6	244366.6	1520269H1	2766	2917
6	244366.6	g4095144	2611	2917
6	244366.6	1237708H1	2767	3016
6	244366.6	2768411H1	2770	3013
6	244366.6	g1321416	2615	2858
6	244366.6	2416809H1	2800	2917
6	244366.6	1570846H1	2811	3018
6	244366.6	3898114H1	2621	2859
6	244366.6	874422H1	2824	3131
6	244366.6	g1349372	2825	2952
6	244366.6	2815235H1	2842	3116
6	244366.6	4897079H1	2925	3201
6	244366.6	1231478H1	2632	2861
6	244366.6	2152887H1	2927	3042
6	244366.6	3510024H1	2635	2917
6	244366.6	1975441F6	2938	3306
6	244366.6	2189605H1	2938	3203
6	244366.6	1975441H1	2938	3088
6	244366.6	3470739H1	2938	3153
6	244366.6	g1844965	2643	2917
6	244366.6	1623474H1	2155	2382
6	244366.6	1338343H1	1799	2055
6	244366.6	2022520H1	2157	2423
6	244366.6	805131H1	2200	2397
6	244366.6	795024H1	2202	2393
6	244366.6	1338343F6	1799	2241
6	244366.6	1297158H1	1810	2050
6	244366.6	3354386H1	2223	2491
6	244366.6	2540345H1	2224	2461
6	244366.6	2313137H1	1887	2152
6	244366.6	2805024H1	2233	2536
6	244366.6	2454581T6	2281	2827
6	244366.6	g570404	1961	2245
6	244366.6	2773281H1	2282	2528
6	244366.6	3246058T6	2284	2807
6	244366.6	3332425T6	2286	2817
6	244366.6	3321733H1	1988	2109
6	244366.6	g2153937	2324	2754
6	244366.6	1464642H1	1995	2224
6	244366.6	g1319564	2331	2938
6	244366.6	3555988H1	2005	2304
6	244366.6	3384030H1	2043	2317
6	244366.6	g1898453	2332	2760
6	244366.6	g1062443	2337	2746
6	244366.6	3188982H1	2348	2686

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
6	244366.6	2255759H1	2045	2318
6	244366.6	2112529H1	2355	2621
6	244366.6	3693731H1	2364	2662
6	244366.6	1864803H1	2069	2351
6	244366.6	853003T6	2385	2815
6	244366.6	g1925211	2137	2615
6	244366.6	2641982H1	2385	2598
6	244366.6	2516658H1	2397	2535
6	244366.6	1864803T6	2414	2898
6	244366.6	1338343T6	2421	2814
6	244366.6	1398258H1	2440	2681
6	244366.6	1829874H1	2471	2738
6	244366.6	589137H1	2478	2685
6	244366.6	4855638H1	2147	2411
6	244366.6	1698592H1	2500	2726
6	244366.6	3524562H1	2154	2348
6	244366.6	g1319504	2523	2952
6	244366.6	1623474F6	2155	2682
7	405313.4	4640462H1	573	837
7	405313.4	g1774849	595	979
7	405313.4	4721077H1	54	194
7	405313.4	5944975H1	61	370
7	405313.4	1948647H1	596	828
7	405313.4	1592016H1	86	282
7	405313.4	1948647R6	596	1136
7	405313.4	g4070751	686	1137
7	405313.4	2384959H1	86	263
7	405313.4	4571373H1	715	978
7	405313.4	g954058	893	1203
7	405313.4	1559555H1	903	1120
7	405313.4	1559555F6	903	1363
7	405313.4	g617633	914	1316
7	405313.4	1302977H1	86	257
7	405313.4	4215272H1	919	1195
7	405313.4	g1492868	88	230
7	405313.4	4643815H1	962	1212
7	405313.4	965308H1	968	1255
7	405313.4	965308R1	968	1622
7	405313.4	1321926T6	132	483
7	405313.4	5136028H1	993	1266
7	405313.4	g4264253	155	609
7	405313.4	g3739298	156	610
7	405313.4	4306178H1	1008	1207
7	405313.4	g1237752	1009	1175
7	405313.4	4551446H1	1047	1310
7	405313.4	g4522654	210	522
7	405313.4	1628853H1	1049	1219
7	405313.4	1627193H1	1049	1261
7	405313.4	1316291H1	349	522
7	405313.4	1628853F6	1049	1649

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
7	405313.4	1283759H1	1053	1330
7	405313.4	4312209H1	1067	1367
7	405313.4	4103086H1	1115	1239
7	405313.4	g1984348	1147	1472
7	405313.4	g2540589	1168	1616
7	405313.4	g1492809	369	527
7	405313.4	3584223H1	1177	1354
7	405313.4	102569H1	458	607
7	405313.4	4829437H1	540	737
7	405313.4	3674811H1	1186	1495
7	405313.4	1733325H1	1192	1412
7	405313.4	g1984560	561	804
7	405313.4	g2010449	2002	2335
7	405313.4	2682711H1	2014	2309
7	405313.4	g4535531	2016	2337
7	405313.4	g1773873	2022	2340
7	405313.4	g4137010	2025	2346
7	405313.4	4111488H1	2026	2288
7	405313.4	2153069H1	2034	2315
7	405313.4	870657R1	2041	2613
7	405313.4	870657H1	2041	2250
7	405313.4	5433876H1	2063	2317
7	405313.4	g3162264	2073	2338
7	405313.4	g3057393	2080	2337
7	405313.4	g3872586	2081	2335
7	405313.4	2081907T6	2084	2288
7	405313.4	659926H1	2097	2337
7	405313.4	056422H1	2097	2317
7	405313.4	3486469H1	2105	2337
7	405313.4	817086R1	2127	2337
7	405313.4	817086H1	2127	2388
7	405313.4	817086T1	2127	2280
7	405313.4	g3597649	2142	2335
7	405313.4	3988965H1	2205	2499
7	405313.4	2664989H1	1633	1850
7	405313.4	3962192H1	1645	1776
7	405313.4	2290557H1	1645	1921
7	405313.4	4115475H1	1647	1861
7	405313.4	g670108	1649	1960
7	405313.4	g570685	1650	1964
7	405313.4	2213032H1	1653	1919
7	405313.4	1667188H1	1653	1775
7	405313.4	2285586H1	1653	1841
7	405313.4	990792H1	1669	1968
7	405313.4	1283707T6	1682	2306
7	405313.4	3781966H1	1686	2020
7	405313.4	2402093H1	1700	1946
7	405313.4	3666248H1	1707	1866
7	405313.4	1948647T6	1711	2289
7	405313.4	5681549H1	1746	2012

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
7	405313.4	2885455T6	1752	2292
7	405313.4	1301039H1	1790	2095
7	405313.4	1669180T6	1799	2303
7	405313.4	1669180H1	1808	2038
7	405313.4	3436329H1	1815	1956
7	405313.4	g1624740	1828	2218
7	405313.4	1559555T6	1831	2302
7	405313.4	3106278H1	1846	2121
7	405313.4	2323323T6	1848	2308
7	405313.4	1914641H1	1848	2062
7	405313.4	1628853T6	1854	2300
7	405313.4	g2946487	1855	2335
7	405313.4	2081907F6	1854	2313
7	405313.4	2081917H1	1854	2015
7	405313.4	g672957	1859	2199
7	405313.4	2916179H1	1878	2177
7	405313.4	3556793T6	1881	2306
7	405313.4	g2783325	1885	2345
7	405313.4	1268187F1	1888	2344
7	405313.4	1268187H1	1888	2152
7	405313.4	1268187F6	1888	2203
7	405313.4	1268187T6	1890	2317
7	405313.4	g616527	1892	2244
7	405313.4	g3593850	1896	2335
7	405313.4	g573001	1897	2264
7	405313.4	g815353	1898	2253
7	405313.4	g4083770	1902	2335
7	405313.4	g4281927	1912	2337
7	405313.4	g2753877	1930	2345
7	405313.4	2199211H1	1930	2188
7	405313.4	2375908H1	1934	2181
7	405313.4	g3797979	1954	2337
7	405313.4	2653312H1	1963	2221
7	405313.4	g4265408	1967	2342
7	405313.4	g3919084	1969	2335
7	405313.4	g4085496	1970	2334
7	405313.4	g668546	1976	2156
7	405313.4	2149388H1	1985	2274
7	405313.4	2601556H1	1995	2280
7	405313.4	g2789326	2898	3179
7	405313.4	g646138	2913	3179
7	405313.4	g888680	2916	3206
7	405313.4	g646137	2924	3179
7	405313.4	g645108	2927	3179
7	405313.4	g917579	2933	3178
7	405313.4	g3051580	2943	3179
7	405313.4	g3764150	2943	3179
7	405313.4	g2903435	2947	3179
7	405313.4	g1225232	2947	3179
7	405313.4	g1087854	2947	3111

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
7	405313.4	g1647814	2998	3213
7	405313.4	g917686	3058	3210
7	405313.4	1226453T1	3087	3168
7	405313.4	1226453H1	3087	3167
7	405313.4	g3174481	3089	3178
7	405313.4	g646728	2878	3179
7	405313.4	g884856	2884	3211
7	405313.4	g815354	2895	3216
7	405313.4	2714769H1	2231	2345
7	405313.4	1440914H1	2252	2421
7	405313.4	1440914F6	2252	2675
7	405313.4	4311940H1	2263	2546
7	405313.4	6073069H1	2266	2498
7	405313.4	549605H1	2285	2554
7	405313.4	2157285H1	2287	2523
7	405313.4	2323179H1	2324	2460
7	405313.4	2323108H1	2324	2581
7	405313.4	g2197270	2337	2696
7	405313.4	2294826H1	2444	2517
7	405313.4	g1043992	2444	2683
7	405313.4	3860110H1	2444	2682
7	405313.4	3246952H1	2472	2725
7	405313.4	1956178H1	2480	2773
7	405313.4	4312886H1	2495	2794
7	405313.4	g770052	2623	2930
7	405313.4	1415784H1	2660	2922
7	405313.4	g884855	2666	3025
7	405313.4	2014171H1	2668	2943
7	405313.4	g888679	2667	3021
7	405313.4	2224791H1	2686	2955
7	405313.4	4106915H1	2695	2996
7	405313.4	g2217789	2772	3181
7	405313.4	g2874275	2772	3179
7	405313.4	1440914R1	2773	3179
7	405313.4	g4075424	2775	3179
7	405313.4	g4328896	2776	3179
7	405313.4	287748H1	2778	3143
7	405313.4	3496950H1	2802	3087
7	405313.4	2750652H1	2806	3105
7	405313.4	g765774	2820	3182
7	405313.4	g3895056	2831	3179
7	405313.4	g4450984	2833	3179
7	405313.4	g2099917	2853	3348
7	405313.4	g2018248	2860	2990
7	405313.4	g564562	2869	3179
7	405313.4	g2459191	2869	3206
7	405313.4	g1099005	2521	2798
7	405313.4	2731146H1	2559	2794
7	405313.4	g1198836	2578	2840
7	405313.4	2229784H1	2603	2852

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
7	405313.4	2229548H1	2603	2858
7	405313.4	1321926H1	1	235
7	405313.4	1321926F6	1	388
7	405313.4	1337104H1	8	265
7	405313.4	2682113H1	8	281
7	405313.4	g2754255	1191	1654
7	405313.4	4541436H1	1206	1470
7	405313.4	g2558152	1209	1673
7	405313.4	g2540638	1218	1616
7	405313.4	1282095H1	1225	1484
7	405313.4	1283707H1	1225	1511
7	405313.4	1282048H1	1225	1506
7	405313.4	1283707F6	1226	1659
7	405313.4	3280293H1	1234	1511
7	405313.4	5398765H1	1241	1391
7	405313.4	5072115H1	1293	1556
7	405313.4	5219645H1	1319	1443
7	405313.4	1893001H1	1317	1616
7	405313.4	3696826H1	1319	1614
7	405313.4	5920192H1	1319	1647
7	405313.4	g3770003	1322	1609
7	405313.4	5665259H1	1327	1555
7	405313.4	3151495H1	1326	1625
7	405313.4	3357256H2	1345	1500
7	405313.4	692886H1	1353	1604
7	405313.4	5400503H1	1388	1623
7	405313.4	2323323H1	1395	1657
7	405313.4	2323323R6	1395	1895
7	405313.4	4622312H1	1404	1718
7	405313.4	5373496H1	1438	1700
7	405313.4	624120H1	1456	1728
7	405313.4	2535460H1	1469	1750
7	405313.4	2115405H1	1505	1801
7	405313.4	g954059	1516	1748
7	405313.4	2292561H1	1538	1790
7	405313.4	2556287H1	1580	1842
7	405313.4	g866975	1610	1953
7	405313.4	861911R1	1619	2200
7	405313.4	861911H1	1619	1876
7	405313.4	g873285	1626	2006
8	436857.2	232864F1	1467	1959
8	436857.2	2704880T6	1506	1925
8	436857.2	g4373224	1522	1968
8	436857.2	g4112872	1522	1954
8	436857.2	g4390509	1556	1962
8	436857.2	5858881H1	1624	1888
8	436857.2	5267222H1	1653	1883
8	436857.2	1477850T6	1680	2225
8	436857.2	4617960T6	1761	2215
8	436857.2	g917596	1901	2223

TABLE 4

SEQ ID NO:	Template ID	Component ID	Start	Stop
8	436857.2	3270104H1	1906	2161
8	436857.2	5782044H1	2003	2209
8	436857.2	g4268407	2010	2240
8	436857.2	g3051680	2063	2241
8	436857.2	481468H1	2063	2250
8	436857.2	5683178H1	1	212
8	436857.2	1477850H1	56	278
8	436857.2	1477850F6	56	488
8	436857.2	4619212H1	138	290
8	436857.2	g1978924	207	556
8	436857.2	3758509H1	206	498
8	436857.2	4617960H1	323	550
8	436857.2	4718961H1	323	555
8	436857.2	4617960F6	323	761
8	436857.2	1992924H1	366	651
8	436857.2	g4269060	528	627
8	436857.2	4255690H1	599	830
8	436857.2	4761770H1	727	1004
8	436857.2	4613106H1	795	1034
8	436857.2	g2000739	795	1025
8	436857.2	2704880H1	883	1176
8	436857.2	2704880F6	883	1313
8	436857.2	2707669H1	961	1264
8	436857.2	805609H1	1054	1283
8	436857.2	805609R1	1054	1630
8	436857.2	4135963H1	1113	1415
8	436857.2	4294249H1	1152	1398
8	436857.2	5450335H1	1156	1421
8	436857.2	5373082H1	1203	1419
8	436857.2	4190554H1	1247	1412